Video Representation Using Greedy Approximations Over Redundant Parametric Dictionaries

Oscar Divorra Escoda, Pierre Vandergheynst and Michel Bierlaire
{oscar.divorra,pierre.vandergheynst,michel.bierlaire}@epfl.ch
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute (ITS)
CH-1015 Lausanne, Switzerland
Technical Report ITS-2004.019

Abstract

In this work, we explore a framework for the sparse representation of video sequences by means of spatio-temporal functions able to exploit the 2D nature of images as well as the temporal smoothness often associated to object trajectories. Decomposition over redundant dictionaries formed by 2D functions capable to exploit image geometry, or more precisely contours orientation, has shown to be well adapted for efficient sparse image approximations. Video representation by means of temporally evolving sets of such 2D functions seems thus a natural extension toward video approximation techniques. In the present paper we study the deformation of a geometry oriented image expansion based on Matching Pursuits (MP) [2], to obtain a parametric representation of frames transformation through time. We consider a modified MP approach based on Bayesian decision criteria to deform geometrical primitives in a predictive fashion from frame to frame. Indeed, since motion stability is not guaranteed using a pure MP, a Bayesian framework is introduced to regularize motion among expansion terms of frames representations.

Index Terms

Video Representation, Spatio-Temporal Approximations, Matching Pursuits, Redundant Dictionaries, Over-complete Expansions, Sparse Representations, Wavelets

Web page: http://lts2www.epfl.ch

Part of this work appeared in ICIP03 [1]. This work has been supported by the Swiss Federal Office for Education and Technology under grant number 6044.1 KTS.

Oscar Divorra Escoda and Prof. Pierre Vandergheynst are with the Signal Processing Institute (ITS) at the Swiss Federal Institute of Technology in Lausanne (EPFL). Web page: http://lts2www.epfl.ch.

Michel Bierlaire is with the Mathematics Institute (IMA) at the Swiss Federal Institute of Technology in Lausanne (EPFL)

Contents

I. Introduction

Video representations are often based on the assumption that, up to occlusions, objects and regions follow smooth geometrical transformations through time [3]. In this paper we consider a basic model of moving pictures where each frame can be represented by a mixture of homogeneous regions and regular contours. Motion is represented in this framework by smooth local deformations of those regions. Such a piecewise smooth model for each frame emphasizes the need for geometry aware representations [4], [5], [6], [7]. But coping with smooth geometric deformations necessitates the use of flexible visual primitives. In order to achieve this we advocate the use of parametrized over-complete dictionaries of basic waveforms, referred to as atoms, along the lines of [8]. Local deformations are propagated along the sequence by updating the atoms' parameter field in order to approximate the succession of frames. Decomposing an image over a redundant dictionary in a stable way being a challenge in itself, we use a simple greedy algorithm, also known as Matching Pursuit (MP) [2], [9] . We consider a modified MP approach based on Bayesian decision criteria to project geometrical primitives through time. A Markov Random Field (MRF) framework is introduced to regularize the obtained motion field. Finally, results are presented to illustrate the effects of the redundant representation and of the regularization technique.

The outline of the paper is as follows. In section II image and sequence models are discussed. In section III, a short review on the expansion of images over redundant parametric dictionaries is found and in section IV we describe the prediction strategies investigated in this work. Greedy stability properties are considered and discussed in section V. In section VI and VII we introduce the use of regularity, suggested by the previous section, and a rate-distortion formulation. Finally, results on the successive approximation of frames are presented in section IX, followed by some discussions and conclusions in sections X and XI.

II. Signal Models

A. Image Model

In this paper we will model images as very short (i.e. sparse) linear superpositions of atoms taken out from a huge, usually very redundant, library of functions $\mathcal{D}$ usually referred to as a dictionary:

$$f_\Gamma(\mathbf{x}) = \sum_{n=0}^{N-1} c_n g_n(\mathbf{x}) \, , \, g_n \in \mathcal{D}, \tag{1}$$

where $\mathbf{x}$ denotes the plane coordinates vector. The efficiency of this model relies on the assumption that any image $f \in L^2(\mathbb{R})$ can be sufficiently well approximated by a given N-term approximant f_Γ. The main motivation for this methodology comes from the complexity of natural images. Common models, such as those underlying wavelet techniques, impose strong assumptions on the signal: usually that the latter is piecewise smooth. Real-world scenes on the other hand cannot be efficiently captured by such limited frameworks. Indeed most images are composed of smooth regions, but the geometrical arrangements of their edges is also very important, and most natural images also contain textures. It has been shown over the past five years that wavelets fail to capture the geometrical information of edges and that this failure turns into non-optimal approximation rates and rate-distortion behavior [5], [4], [7]. In order to adapt to more complex image models one of us has already advocated the use of redundant libraries of basic image primitives [2], [10]. The main advantage of this method compared to more classical representations using more rigid ortho basis or frames lies in the freedom one has to design the atoms. They can be tailored to efficiently capture prominent image structures such as edges or textures. They can also be used to approximate mixtures of models. This freedom comes at a price though: without more restriction on the dictionary, no fast algorithm will usually exist. Moreover, and more disturbing, there usually does not exist a constructive algorithm to manipulate these libraries. By this we mean the following : suppose there exists a unique optimal f_Γ in the sense that

$$f_\Gamma = \underset{\substack{\Gamma, |\Gamma| \leq N \\ \mathbf{c} \in \mathbb{R}^{|\Gamma|}}}{\arg\min} \left\| f - \sum_{k \in \Gamma} c_k g_k \right\|_2^2 , \tag{2}$$

where $\Gamma \subset \Omega$ such that $\mathcal{D} = \{g_\gamma : \gamma \in \Omega\}$. Then finding f_Γ is usually a NP-hard problem.

Hope can nevertheless be found in recent advances in redundant approximations [11], [12]. It is known, for example, that if $\mathcal{D}$ is not too redundant, problem (2) can be solved in polynomial time. More recently, [13] has shown that it is also possible to use non-orthogonal greedy heuristics to solve this problem in a stable manner: only atoms belonging to the optimal set Γ would be recovered by a greedy algorithm. However, very redundant over-complete dictionaries are often used in practice for the approximation of images. In that case, the recovery of the optimal set Γ will not be guaranteed. Anyway, sub-optimal solutions to (2) can still be good sparse approximations of f. Highly redundant dictionaries can be composed by functions that closely match local signal structures, which contributes to a good approximation error decay with few terms. We will come back to the stability of (weak) greedy algorithms later in this paper.

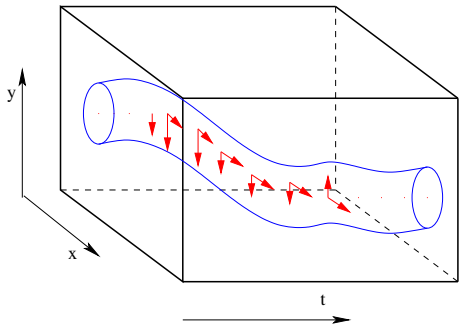

Fig. 1. Schematic smooth evolution of an object through time.

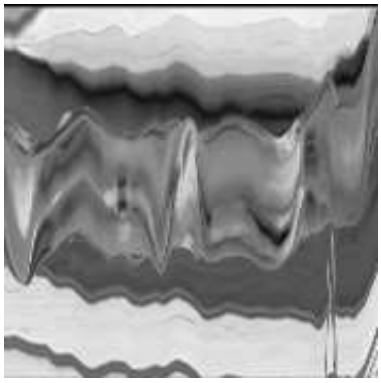

Fig. 2. Temporal evolution of a pixels row (the 77th from QCIF version) in foreman (from frame 0, at top left, until frame 176).

B. Sequence Model

Natural image sequences are composed of successive projected snapshots of 3D objects. Considering these objects to be smooth and their trajectories to be smooth functions of time, one usually assumes that image sequences are well modeled by smooth transformations of a reference frame [3]. Of course this assumption has intrinsic limitations : natural sequences display a wide variety of transient behaviors such as occlusions, appearance and disappearance ... A schematic illustration is depicted in Fig. 1. This basic smoothness assumption is also "visually" justified in Fig. 2 where we display a section (a line here) of the Foreman sequence. Time is on the horizontal axis and the whole image looks very smooth.

The local geometric transformations mentioned above are tightly linked with the motion model and the nature of the real 3D scene. When the support of moving regions is sufficiently small, a simple translational model can successfully represent motion. This is the key ingredient of most block matching techniques in motion compensation : the reference frame is chopped into small primitives which are assumed to just translate in time. The whole model is then represented by a translation vector field. Complex and more accurate models have also been considered, for example affine models [14]. These allow for local expansions or contractions and are usually represented by mesh deformations [15]. More precisely the motion model is locally represented by a linear transformation :

$$\mathbf{u} = \mathbf{A}\,(\mathbf{x} - \mathbf{b})\,. \tag{3}$$

The 2×2 squared matrix $\mathbf{A}$ and vector $\mathbf{d}$ implement translation, rotation, shearing and scaling operators:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \frac{1}{sx} & 0 \\ 0 & \frac{1}{sy} \end{bmatrix} \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \\ \mathbf{b} &= \begin{bmatrix} b_x \\ b_y \end{bmatrix}, \end{aligned} \tag{4}$$

where sx and sy are scale parameters, m is a sheering factor to change the angle among x and y axis ($m = 0$ if axis are kept perpendicular) and θ parametrizes geometric changes due to rotations.

Most video representation paradigms separate motion information from image structures. In this paper, we would like to jointly represent image geometrical structures *and* their evolution through time. Given a set of images belonging

to a sequence, the changes suffered from frame I_t to I_{t+1} are modeled as the application of an operator F on the image I_t such that

$$\begin{aligned} I_{t+1} &= F_t\left(I_t\right), \\ I_{t+2} &= F_{t+1}\left(I_{t+1}\right) = F_{t+1}\left(F_t\left(I_t\right)\right), \\ I_{t+3} &= ... \end{aligned} \tag{5}$$

where the subindex t corresponds to time.

From Eq. (1) and (5) we thus model $\hat{I}_{t+1}$ as a transformation of the geometric representation of $\hat{I}_t$:

$$\hat{I}_{t+1} = F_t\left(\sum_{\gamma\in\Gamma} c_\gamma^t \cdot g_\gamma^t\right). \tag{6}$$

A relation needs to be established between F_t and the transformation of each one of the 2D components involved in the frame approximation. This is why we make the hypothesis that F_t is composed by the set of F_t^γ that independently transform each one of the frame expansion terms, i.e.

$$\hat{I}_{t+1} = \sum_{\gamma\in\Gamma} F_t^\gamma\left(c_\gamma^t \cdot g_\gamma^t\right). \tag{7}$$

In the following we will sometimes refer to F_t as a *deformation*. The action of each F_t^γ in (5) corresponds to a geometrical operation on g_γ and to a change of its coefficient c_γ^t. Intuitively, this mechanism intends to implement a local change of scale, position and direction of each primitive (see Fig. 4(a) and 4(b)). The sequence of deformed atoms g_γ^t can thus be seen as a 3D primitive representing how local scene geometry flows through time.

III. Expansions Over Redundant Parametric Dictionaries

Eq. (1) sets our general framework. In this equation, the atoms $\{g_\gamma : \gamma \in \Gamma\}$ might be particular orthonormal bases or frames selected for their ease of implementation and good approximation properties [16]. As previously stressed though, the sparsity of these schemes is strongly limited by the reduced capability of generic families, such as wavelets, to accurately model complex signal features. This is the key argument motivating the exploration of redundant dictionaries: the closer the g_γ are to these basic features, the sparser our model (1). In a more mathematical phraseology, a close relationship binds good signal modeling and sparsity through non-linear approximation rates. Unfortunately, estimating such rates is a daunting task when considering very redundant dictionaries and we must rely on heuristics and intuition to design correct atoms.

Another puzzling problem one faces with redundant dictionaries is that in general there is no unique solution to (2). Nevertheless, as stated in the previous section, various algorithms have been introduced to find good approximates. Once again choices have to be made, involving a mix of mathematical modeling and heuristics that we will now discuss.

A. Image Approximation

As introduced in Sec. II-A we assume a piecewise geometry oriented model for image representation. Many dictionaries could be considered for the modeling of edges, from wavelet oriented strategies (where several functions may be needed to approximate a sharp edge) to more *curvelet* like dictionaries (where edges are treated through dedicated atoms). Based on previous considerations [10], [17], [2], we chose an edge oriented dictionary based on rotated and anisotropically scaled functions. This dictionary is based on an edge-detector function, called *Anisotropic Refinement* (AR) Atom described in [8], [18]:

$$g(u,v) = \frac{1}{C}\left(4u^2 - 2\right)\exp\left(-\left(u^2+v^2\right)\right), \tag{8}$$

where C is a normalizing constant that guarantees unit L_2 norm.

In this work, the dictionary is build from a variety of affine transformations of (8), these are based in our case on the Euclidean group and the anisotropic scaling of (u,v) to represent geometric features. Indeed, according to our experiences with images, sheering is not of paramount importance and will not be used here:

$$\begin{aligned} \mathbf{A}' &= \begin{bmatrix} \frac{1}{sx} & 0 \\ 0 & \frac{1}{sy} \end{bmatrix}\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \\ \mathbf{b} &= \begin{bmatrix} b_x \\ b_y \end{bmatrix}. \end{aligned} \tag{9}$$

1) 2D Dictionary Generation: The dictionary $\mathcal{D} = \{g_\gamma : \gamma \in \Omega\}$[1] is generated by applying a discrete set of geometric transformations to the mother function (8) which is defined such that:

$$g_\gamma = g_{b_x,b_y}^{sx,sy,\theta} = g(u,v), \tag{10}$$

where sx, sy, θ, b_x, b_y are respectively the scalings, rotation and positions parameters defined in (9) and

$$\begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{A}' \begin{bmatrix} x - d_x \\ y - d_y \end{bmatrix},$$

where (x, y) are the image 2D coordinates.

In practice sampling the set Ω, of parameter vectors, is performed such that over-completeness of the dictionary is ensured. This was achieved in [8] by ensuring a good covering of the frequency plane. The parameter set should also be chosen dense enough so that the dictionary is almost covariant through continuous changes of position, rotation and scaling. This feature being of high practical importance as explained in Sec. III-B. Nevertheless there is a clear trade-off between dictionary density and computational complexity. Looking for a compromise, a sampling angular step $\Delta\theta = \frac{\pi}{36}$ is considered, all the coordinate points in the image ($b_x, b_y \in \mathbb{N}$) and a resolution of half octave is considered for both scale parameters (sx, sy).

2) Decomposition Strategy, use of Matching Pursuits: The retrieval of the best approximation according to a given constraint can be a very difficult task. In the scope of our study we consider the simple greedy approach known as MP [9] to find an approximate solution to the sparse representation of Eq. (2) .

In this, the signal is iteratively approximated by adding at each iteration a new term to the signal expansion. Taking as initial condition $\mathcal{R}_0 = I$ in the iterative expression:

$$\mathcal{R}_{n+1} f = \mathcal{R}_n f - \langle \mathcal{R}_n f, g_{\gamma_n} \rangle \, g_{\gamma_n}, \tag{11}$$

where $\mathcal{R}_n f$ is the residual at the *nth* iteration, the function g_γ that minimizes the $\|\mathcal{R}_{n+1} f\|_2$ will be selected for g_{γ_n}. This results on the following expansion:

$$I = \sum_{n=0}^{N-1} c_{\gamma_n} \cdot g_{\gamma_n} + \mathcal{R}_N f, \tag{12}$$

where $c_{\gamma_n} = \langle \mathcal{R}_n f, g_{\gamma_n} \rangle$ and has the property:

$$\|I\|^2 = \sum_{n=0}^{N-1} |c_{\gamma_n}|^2 + \|\mathcal{R}_N f\|^2 . \tag{13}$$

This procedure always converges, and even exponentially in finite dimension. MP can be very good at producing good very sparse approximants, but will in general never yield perfect reconstruction. Orthogonal Matching Pursuit (OMP) [19], however, solves this ever lasting expansion problem of MP by orthogonalizing at every iteration the error at the expense of a higher computational complexity. Our motivation for considering the MP paradigm lays mainly in its applicability in terms of complexity for natural and common size images expansions on highly over-complete dictionaries. Furthermore greedy approaches have the advantage of being flexible (e.g if an additional term is needed just an additional iteration has to be computed). In any case, sub-optimality with respect to a global solution is the cost for the simplicity of the algorithm.

B. Video Approximation: Tracking 2D Image Features Through Time

In this work we study the approximation of F_t such that the set of functions g_γ^t and g_γ^{t+1}, at time t and $t+1$ respectively, belong to the dictionary $\mathcal{D}$:

$$\forall \gamma, \ \forall t \quad g_\gamma^t \overset{F_t^\gamma}{\longmapsto} g_{\gamma'}^{t+1} \text{ s.t. } g_\gamma^t, g_{\gamma'}^{t+1} \in \mathcal{D}. \tag{14}$$

This last imposition is considered for several reasons. The fact that g_γ^t and $g_{\gamma'}^{t+1}$ belong to the same dictionary $\mathcal{D}$ allows the use of the same fast atom search used for image approximations [18] to solve F_t^γ. In the case where the parametric description of a sequence is coded, a quantization on the evolution of geometric parameters γ is required. For a better performance, this quantization has to be embedded in the decomposition loop of the greedy algorithm. The quantization is such that pieces of a 3D time evolving feature embedded in a particular frame, belong to the 2D dictionary in use. In effect, if one desires to reconstruct a particular frame, from the spatio-temporal geometric video representation, this may be done using always the same dictionary $\mathcal{D}$.

[1]From now on, Ω not only refers to the selected index set of atoms, but also the set of parameter vectors associated the selected indexes. In effect, $\gamma \in \Omega$ and $\gamma = \{b_x, b_y, sx, sy, \theta\}$.

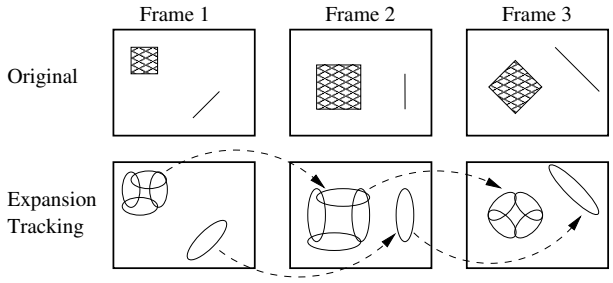

Fig. 3. Successive schematic updates of basis functions in a sequence of frames.

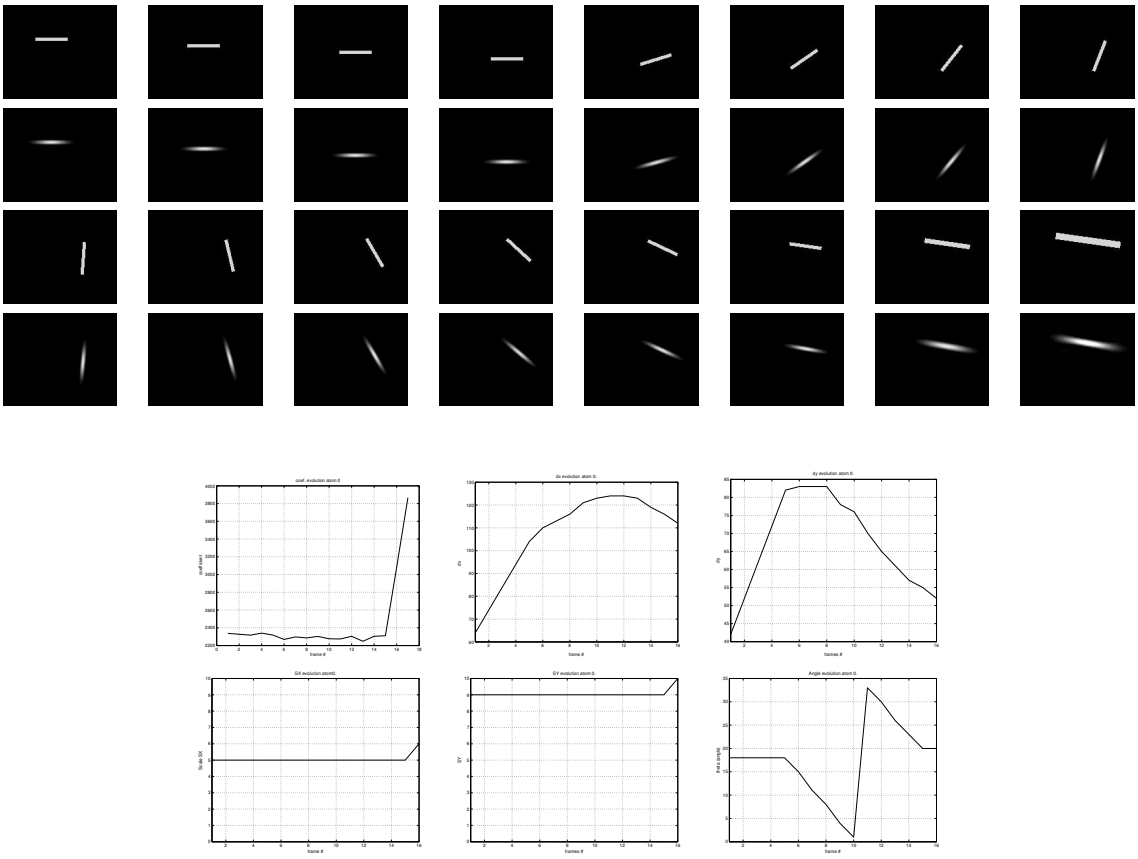

Fig. 4. Up: Synthetic sequence approximated by 1 atom. Down: Parameter evolution of the approximated object; from left to right and from up down, we find: coefficient, x position, y position, x (short axis) scale, y (long axis) scale, angle.

The set of all possible transformations F_t^γ is an approximation of the affine model of local transformations defined for sequences. This approximation intends to supply a trade off between adaptation flexibility and dictionary complexity, i.e. it does not include the model of shearing and is limited by the granularity of the dictionary parameters.

Looking to recover the transformations of geometrical features, we formulate the problem from a frame to frame point of view. Fig. 3 schematically describes the approximation of operator F_t of Eq. (5) studied in this work. A more practical example can be seen in Fig. 4(a) where the approximation of a simple synthetic object by means of a single atom is performed. The first and third row of pictures show the original sequence and the second and fourth rows provide the reconstruction of the approximation. Fig. 4(b) shows the parametric representation of the sequence. We see the temporal evolution of the coefficient c_γ^t, the coordinates evolution of the translation parameters and the scale and angle evolution.

The problem that we are facing (akin to sparse image expansions in sec. III-A.2) may be seen as the optimization

of:

$$\min_{F_t} \left\| I_{t+1} - \sum_{\gamma \in \Gamma} F_t^{\gamma} \left[c_{\gamma}^{t} \cdot g_{\gamma}^{t} \right] \right\|_2 \quad \text{subject to} \quad \text{Cost}\left(F_t\right) \leq \xi, \tag{15}$$

where "Cost" is a constraint depending on the application. This optimization turns out to be very complex and even NP-hard depending on its formulation and the *Cost* measure. However, and depending on the scale of the problem, some case could be considered where a global optimization is feasible.

Example 1: Consider the simplified problem where only the selection of functions is involved (thus, coefficients $c \subset \{0,1\}$) and where the costs for every function can be predefined in a fixed manner in the form of a vector ($\mathbf{w}$ where $\forall i \; w_i \geq 0$) where each w_i determines the selection cost of the i*th* function and the cost constraint can be rewritten as $\mathbf{w}^T\mathbf{c} = Cost$. Thus, Eq. (15) could be redefined as:

$$\min_{{c_{\gamma}}^{t+1} : \gamma \in \Gamma} \left\| I_{t+1} - \sum_{\gamma \in \Gamma} G_{\gamma}^{t+1} {c_{\gamma}}^{t+1} \right\|_2 \quad \text{subject to} \quad \mathbf{w}^T\mathbf{c} \leq \xi, \tag{16}$$

where G_{γ}^{t+1} is the set of functions in which a given g_{γ}^{t} may potentially transform and every selection vector ${c_{\gamma}}^{t+1}$ a boolean vector such that $\forall \gamma \; 0 \leq \|{c_{\gamma}}^{t+1}\|_1 \leq 1$. To be clearer, $\mathbf{c}$ in the cost constraint is the concatenation of all c_{γ}. Looking to Eq. (16) the problem resembles a lot to the formulations of the retrieval of the best m-term approximation, except for the additional linear constraint on the costs, which modifies slightly the constraint norm. More generally, problems of the kind of (16) are instances of the *Knapsack* problem [20].

IV. Tracking Frame Deformations

In order to obtain a parametric representation in terms of the evolution of geometrical components, a greedy approach is considered for the progressive approximation of every video frame. This approach, very close to the one taken for still images, consist in approximating every primitive transformation in a successive manner. However, some further considerations are needed given the assumed motion model. This has the purpose of reducing computational complexity and ensure locality and *smoothness* of the motion parameters. In fact, direct MP full search in a frame at $t+1$ does not have any guarantee to recover the corresponding deformed atoms from frame t. Greedy algorithms are sub-optimal and myopic: they are limited by the resolution of the dictionary [21]. In the case of motion, a simple image deformation may induce MP to select the wrong primitive transformation. Moreover, this can contribute to propagate the primitive selection error to posterior MP iterations.

A. The Weak Greedy Algorithm

Imposing additional constraints to the selection rule of MP [9] can sometimes be modeled by weakening its greedy nature, i.e. select g_{γ_n} such that

$$\left| \left\langle \mathcal{R}_n^t f, g_{\gamma_n} \right\rangle \right| \geq \alpha_n \sup_{\gamma \in \Gamma} \left| \left\langle \mathcal{R}_n^t f, g_{\gamma} \right\rangle \right|, \tag{17}$$

where $\alpha_n \in [0,1]$. This MP variation, studied in detail in [22], is known under the name of *Weak* Greedy Algorithm (WGA), *Weak* MP or *Weak*(α) MP when a the flexibility factor (α) is considered to be independent of n.

B. Greedy Local Search

The assumption of local deformation of the sequence model, imposes that the transformation F_{γ}^{t} of a given atom g_{γ}^{t} can not result in any function $g_{\gamma} \in \mathcal{D}$. As in the case of Block Matching (BM) [3] some constraint on the search space can be set. Solutions beyond the search space are considered to be very improbable. Furthermore, the functional to be optimized is non-convex (15). This may yield a slightly better match away from the appropriate place, breaking consequently the structure of the approximation. A local greedy heuristic is defined by means of a sub-dictionary $\mathcal{D}' \subset \mathcal{D}$ associated to every g_{γ}^{t}. The search for the transformation will be performed in $\mathcal{D}'$ solely. We will consider a range of variations $\Delta\gamma$, i.e. in position ($\pm\Delta b_x, \; \pm\Delta b_y$), scale ($\pm\Delta sx, \; \pm\Delta sy$) and angle ($\pm\Delta\theta$):

$$\mathcal{D}'_{\gamma} = \left\{ g_{\gamma'} : \gamma' \in [\gamma - \Delta\gamma, \gamma + \Delta\gamma] \right\}. \tag{18}$$

Given the analytic expression (8) and the non-convex, non-linear form of the problem may suggest the use of quasi-Newton methods combined with other techniques such line search or trust-region globalization techniques [23]. However, the complex topology of the objective function makes it likely to fall in local minima. Furthermore, the cost of using the analytical expressions is subject to the high computational cost of evaluating exp().

A local full search based on the computation of all the matching positions will avoid local minima at a reasonable cost if properly implemented [18]. As in the case of static images, the use of a precomputed Fourier version of the different atoms generated from Eq. (8), allows the computation of the projection on all tranlated atoms with a single FFT [18].

C. Use of Motion Model Constraints: Multi-objective Optimization

Using a very redundant dictionary (Sec. III-A) improves signal modeling but at the expense of a weaker discrimination between atoms. Some additional information is thus needed in order to select a good candidate. As suggested in [24] for a similar approach for motion estimation, the inclusion of a priori information in the selection functional may help achieving estimates of frame to frame primitive transformations more respectful of the sequence model of Sec. II-B. A possible approach can be to impose a regularity constraint among neighboring primitives. Some interdependence is assumed for primitives belonging to the same structure (Sec. VI). In a coding perspective, however, estimating regularity by means of rate could be more appropriate (Sec. VII).

V. Stability in the MP approximations

The use of a greedy strategy implies that only one of the F_t^γ is optimized at every iteration, without taking into account the possible interdependence this might have with the other $F_t^{\gamma'} : \gamma \neq \gamma'$ operators. If F_t^γ are independent $\forall \gamma \in \Gamma$ (i.e. $g_\gamma^{t+1} \perp g_{\gamma'}^{t+1} \quad \forall \gamma \neq \gamma' \quad \gamma, \gamma' \in \Gamma$), each one of them can be optimized independently, leading the algorithm to work perfectly. However, given the non-orthogonality of our Dictionary, it is not clear whether an MP like algorithm will succeed in giving a good solution to Eq. (15). In some cases the greedy algorithm might wrongly choose a function at a given iteration. In our case, given a frame I at time t and its decomposition of the form (12), we impose that the optimal set of functions that represent the frame at time $t+1$ should be the set of $g_\gamma^{t+1} : \gamma \in \Gamma$ such that F_t^γ approximates the local motion of the scene. We intend thus to recover this *optimal* set of functions to represent the frame at time $t+1$. Considering this, an analogy can be established with the problem of recovering the best m-term sparse representation of a signal [11], [12], [13]. Indeed, a relation between the structure of the dictionary in use and the algorithm used to recover a given sparse superposition of waveforms (12) can be established.

In the following, we first review which conditions ensure that MP will recover a given signal expansion. As detailed below, good MP behavior is constraint to the case of incoherent dictionaries. In order to better match our particular case, an extension of the MP stability result is performed in Sec. V-B and V-C, where additional *a priori* knowledge is considered in the greedy selection criteria. These analysis elucidate how *a priori* information can help the behavior of greedy algorithms when the dictionary incoherence condition is not satisfied.

A. MP Stability

A critical point of greedy algorithms is their stability. This means that given the superposition of functions

$$f = \sum_{\gamma \in \Gamma} c_\gamma g_\gamma \ s.t. \ f \in span\left(g_\gamma, \ \gamma \in \Gamma\right) \ and \ D_\Gamma \subset \mathcal{D}, \tag{19}$$

the MP algorithm using dictionary $\mathcal{D}$ will not necessarily recover the same set of functions Γ. The *exact* recovery of "correct" g_γ will be only ensured if the following *Stability condition* (SC) [13], [11] is satisfied:

$$\sup_{\gamma \notin \Gamma} \left\| \left(D_\Gamma\right)^+ g_\gamma \right\|_1 < 1, \tag{20}$$

where $(\cdot)^+$ denotes the *Moore-Penrose Pseudoinverse*. In the case of *Weak*(α) MP (see 17), α would substitute the bound of 1 in (20). This bound is indicative of the behavior of MP with an over-complete dictionary (e.g $\mathcal{D}$ in sec. III-A.1) and the function (f). Eq. (20) implies that optimal functions that expand the space of f must be different *enough* from any other function of the dictionary, not included in Γ, to be recovered by MP.

Given that optimal atoms are not usually known in advance, an upper bound of Eq. (20) has been established based on a internal coherence of the dictionary in use. As demonstrated in [13], [11], [12], (20) implies that given the following measure of the internal coherence of the dictionary (the *Babel function*), where

$$\mu_1(m) \triangleq \max_{|\Gamma|} \max_{\|\Gamma\|_0 = m} \max_{\gamma \notin \Gamma} \sum_{l \in \Gamma} |\langle g_l, g_\gamma \rangle| \tag{21}$$

then it can be stated:

Theorem 1: (*Gribonval, Vandergheynst*) Let m be an integer such that

$$\mu_1(m) + \mu_1(m-1) < 1. \tag{22}$$

Then for any index set Γ of size at most m, any $f \in span\left(g_{\gamma, \gamma \in \Gamma}\right)$, and $\alpha > \mu_1(m)/(1-\mu_1(m-1))$, *Weak*($\alpha$) General MP picks up a "correct" atom at each step. That is, the *Weak*(α) MP algorithm will retrieve the atoms from the set Γ defined in (19) when (22) is in force.

This result shows that MP will behave well with incoherent dictionaries. It has to be noted that even for very coherent dictionaries, experience shows that MP may behave well. Theorem 1 provides indeed a pessimistic bound for the worst case. We will now refine this theoretical analysis to match our particular algorithm.

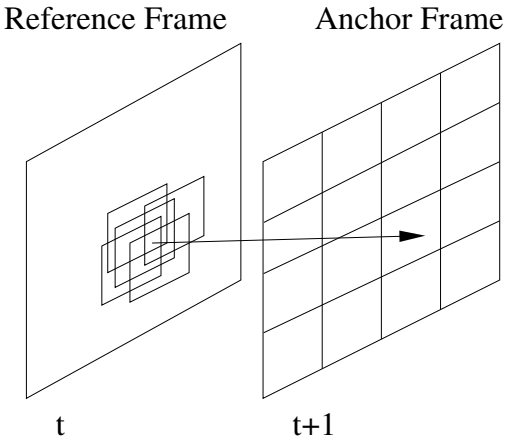

Fig. 5. BM using a fixed block size anchor frame. Each set of candidate blocks to match into a block of the anchor frame are, according to Sec. V-B, an orthogonal dictionary block.

B. Considering the Block Structure of the problem

As described in Sec. IV-B the dictionary used for the prediction of video frames is composed by blocks of candidate functions. Each block has been generated by all admissible transformations of a given primitive from the previous frame. In the approximation of future frames, according to the assumed motion model, only one of these elements will be taken into account, i.e. just an atom from a single block. In this situation and in a way similar to [25], some constraints exist that can be exploited in the definition of an upper bound for the stability condition.

Theorem 2: Let $\mu_{1_B}(m)$ be the inter-block coherence defined in (see Appendix A) for a block dictionary $\mathcal{D} = \bigcup_l \mathcal{D}_{B_l}$ and let μ_{D_B} be the biggest possible inner product among two different functions into a block. If the signal f is such that

$$f \in span\left(g_{\gamma_{B_l}} : l \in [0, m-1],\ g_{\gamma_{B_l}} \in \mathcal{D}_{B_l}\right), \tag{23}$$

(i.e., f belongs to the space generated by a set of m atoms each of them belonging to a different dictionary block) and

$$\frac{\mu_{D_B} + \mu_{1_B}(m-1)}{1 - \mu_{1_B}(m-1)} < \alpha, \tag{24}$$

the Weak(α) algorithm will recover the set of correct atoms that compose f (see App. A for a proof).
Thus, for $\alpha = 1$, we need $\mu_{D_B} + 2\mu_{1_B}(m-1) < 1$. This result implies that in order to recover the "correct" atoms, μ_{D_B} and $\mu_{1_B}(m-1)$ can not be *big* at the same time. A very redundant dictionary (μ_{D_B} close to 1) will need very incoherent blocks in order to ensure the right selection of functions. In the other way round, if blocks are coherent enough, μ_{D_B} needs to be sufficiently small to ensure the exact recovery for any kind of signal.

Example 2: Motion estimation by means of block matching (BM) [3] can be seen as a particular case of the case discussed in here. Consider a correlation based approach where all matching candidates for a given image block are normalized and have zero mean [26]. The anchor frame is divided in non-overlapping blocks. Each of these blocks has to be approximated by the most similar block selected from a set of blocks in the reference frame. This set of blocks correspond to all the possible blocks that belong to a neighboring area (see fig. 5). They would correspond to one of the dictionary blocks described above. Furthermore, since anchor frame blocks do not overlap, dictionary blocks are orthogonal. Thus, as long as there are no identical pixmap pieces into a given dictionary block (i.e., $\mu_{D_B} < 1$), the recovery of the optimal anchor frame expansion is ensured by Theorem 2.

C. Including a Priori Information in the MP Selection Criteria

As already noticed, the use of a redundant dictionary, does not imply that a given algorithm (greedy in our case) will definitely fail recovering appropriate primitives. In fact, this will depend as well on the signal to recover. If *a priori* knowledge on the signal to approximate is available, then further discrimination is possible in addition to the scalar product between signal and atoms.

Let us examine how this can be done.

Definition 1: Let $W(f)$ be a square diagonal matrix of the size of the dictionary $\mathcal{D}$ with $w_{ii} \in [0, 1]$ where each of the w_{ii} corresponds to the *a priori* likelihood of a particular atom g_i to be part of a given signal f.

Definition 2: Let $\mu_1^w(m,f)$ be the following data dependent coherence measure:

$$\mu_1^w(m,f) \triangleq \max_{\Gamma|\ \|\Gamma\|_0=m} \max_{\substack{g_{\overline{\gamma}} \cdot w^{\overline{\Gamma}}_{\overline{\gamma},\overline{\gamma}}, \\ \overline{\gamma} \notin \Gamma}} \sum_{\gamma\in\Gamma} | < g_\gamma, g_{\overline{\gamma}} > | \cdot w^{\Gamma}_{\gamma,\gamma} \cdot w^{\overline{\Gamma}}_{\overline{\gamma},\overline{\gamma}}. \tag{25}$$

See App. B for a detailed description. Note that $\mu_1^w(m,f)$ is defined similarly as in Sec. V-A and each inner product in the sum is weighted by the corresponding probability w_{ii}.

Theorem 3: Given the *a priori* matrix $W(f)$ and weighted coherence $\mu_1^w(m,f)$, the associated Stability Condition is

$$\sup_{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}} \left\| \left(D_\Gamma W_\Gamma\right)^+ g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}} \right\|_1 < \alpha. \tag{26}$$

If $W(f)$ is a *reliable* (in the sense of App. B) *a priori* information, Stability is ensured when

$$\frac{\mu_1^w(m)}{1-\mu_1^w(m-1)} < \alpha. \tag{27}$$

Since $\mu_1^w(m) \leq \mu_1(m)$ then one can intuitively see that a reliable prior knowledge can help a greedy algorithm when the dictionary does not satisfy the hypothesis of Theorem 1.

In the following, we will model *a priori* information using a Bayesian formulation and use it to fine tune our greedy algorithm.

VI. Using Regularity Constraints: A Bayesian Approach of the Problem

The use of a greedy approach to solve the deformation of primitives $g^t_{\gamma_n}$ from frame I^t to approximate I^{t+1} has the drawback of the instability associated to using MP with very redundant dictionaries (as stated in Sec. V). Instability can be further amplified by the sampling of the parameter space that characterizes primitives g_γ. First the search space for geometric primitives adaption is limited to a discrete subset. Moreover the assumption that motion is uniform over the support of a given atom may fail. The use of the simple matching rule of the pure greedy algorithm does not necessarily respect the smooth motion of the sequence: a good match can sometimes be found far from the trajectory of the primitive. The inclusion of an *a priori* model in the greedy selection criteria is thus necessary to reduce instability on the recovery of primitives. A first solution is to perform a local search on a reduced subspace. However more complex models can be taken into account. In this section we show how Bayesian modeling can be used to tackle that problem.

A. Probability Model to Optimize

We reformulate the greedy selection criteria from a probabilistic point of view. Intuitively taking the strongest scalar product (Eq. (17)) consists in selecting the most probable atom. However, much more involved probabilistic models can be considered. The probability space can involve other variables. A Bayesian modeling of the problem can be performed if some a priori information or knowledge about the parametric sequence description is available. We make the assumption that in the sequence approximation (Eq. (15)) neighboring atoms present regular motion since several of them are needed to represent a region. This regularity has the role of the *Cost* term in (Eq. (15)). In the greedy formulation, a Bayesian functional that maximizes the Maximum a Posteriori (MAP) probability will integrate the regularity motion assumption. We consider a Markov Random Field (MRF) framework to define probabilistic relations among atoms.

Thus, for every MP iteration we optimize:

$$\begin{aligned} \Delta\gamma_n = \arg\max_{\Delta\gamma_n} \left\{ p\left(\Delta\gamma_n, \Delta c_n \mid \mathcal{R}_n^{t+1} f, g^t_{\gamma_n}\right)\right\} = \\ \arg\max_{\Delta\gamma_n} \left\{ p\left(\mathcal{R}_n^{t+1} f, g^t_{\gamma_n} \mid \Delta\gamma_n, \Delta c_n\right) \cdot \right. \\ \left. p\left(\Delta\gamma_n, \Delta c_n\right)\right\}, \end{aligned} \tag{28}$$

such that $\Delta\gamma_n = \gamma_n^{t+1} - \gamma_n^t$ and $\gamma_n^{t+1}, \gamma_n^t \in \Gamma$. In Eq. (28) the most probable transformation is taken given a residual at $t+1$ and the corresponding g_γ at time t for a given greedy step n. By the Bayes' rule, this is equivalent to maximizing the probability of matching a given $g^t_{\gamma_n}$ with the residual $\mathcal{R}_n^{t+1}$ conditioned to the probability of the transformation $\Delta\gamma_n$ and the temporal change on the projection coefficient Δc_n. The matching probability $p\left(\mathcal{R}_n^{t+1} f, g^t_{\gamma_n} \mid \Delta\gamma_n, \Delta c_n\right)$ can be defined as a function of an estimated residual error energy $\left\|\hat{\mathcal{R}}_{n+1}^{t+1} f\right\|^2$ for the retrieval of function g_{γ_n} at iteration n. Atoms are assumed to deform under consistent motion transformation. Thus, no change in the coefficient will be considered (except for scale changes) in the estimation of the most probable motion:

$$\hat{\mathcal{R}}_{n+1}^{t+1} f = \mathcal{R}_n f^{t+1} - \overline{\langle \mathcal{R}_n^t f, g^t_{\gamma_n} \rangle} g^{t+1}_{\gamma_n}, \tag{29}$$

where $\overline{\left\langle \mathcal{R}_n^t f, g_{\gamma_n}^t \right\rangle}$ is normalized according to a possible re-scaling of $g_{\gamma_n}^{t+1}$ with respect to $g_{\gamma_n}^t$, i.e $\overline{\left\langle \mathcal{R}_n^t f, g_{\gamma_n}^t \right\rangle} = \left\langle \mathcal{R}_n^t f, g_{\gamma_n}^t \right\rangle \sqrt{\Delta sx \Delta sy}$. At time t, $\mathcal{R}_{n+1}^t f \perp g_{\gamma_n}^t$, in order to minimize the energy of the projection error. In the same way, after motion transformation, $g_{\gamma_n}^{t+1}$ should be such that $\left\| \hat{\mathcal{R}}_{n+1}^{t+1} \right\|^2$ is also minimized.

The probability measure assumes the Gaussianity (by the central limit theorem [27]) and independence of error samples $\mathcal{R}_{n+1}^t f(x,y)$ (although this is not often the case for this class of signals). Based on previous approaches of the block matching and MRF fields [28], [29], we consider:

$$p\left(\mathcal{R}_n^{t+1} f, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) = \frac{1}{Z} \prod_{x,y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|\hat{\mathcal{R}}_{n+1}^{t+1} f(x,y)|^2}{2\sigma^2} \right) \tag{30}$$

where Z is a normalizing constant and $\sigma^2 \approx E\left[\left| \hat{\mathcal{R}}_{n+1}^{t+1} f(x,y) \right|^2 \right]$. Note that $\hat{\mathcal{R}}_{n+1}^{t+1} f$ is considered to have zero mean. In fact, prior to any operation, a low pass approximation is removed from every frame (see Sec. VIII). Introducing the evaluation of σ^2 in Eq. (30) we obtain the conditioned optimization criteria:

$$p\left(\mathcal{R}_n^{t+1} f, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) \approx \frac{C_1}{\sqrt{\left\| \hat{\mathcal{R}}_{n+1}^{t+1} \right\|^2}}, \tag{31}$$

where C_1 is a constant.

The probability $p\left(\Delta\gamma_n, \Delta c_n\right)$ imposes the model that drives the transformation F_t^γ of the g_γ^t and the associated coefficient. It is thus defined as the conditioned probabilities of the $\Delta\gamma$ and Δc_n in the framework of MRFs. At every iteration, MP will try to select a new atom that maintains regularity with all previously selected primitives in the neighborhood. Earlier atoms are trusted to generate the MRF for the future appearing atoms. This unbalanced criteria derives from the fact that first atoms of the MP decomposition capture more energy, thus they tend to represent much more significant (i.e. reliable) features from the signal. When no first reliable estimate of $p\left(\Delta\gamma_n\right)$ is available , an initial tentative needs to be performed trying to match the whole region where the atom is supported. Atoms interaction, MP sub-optimality and simplicity of atoms waveform may reduce the reliability of the estimation provided by a single atom matching. This will be explained later in Sec. VIII.

We can formulate $p\left(\Delta\gamma_n, \Delta c_n\right)$ as:

$$p\left(\Delta\gamma_n, \Delta c_n\right) = p\left(\Delta c_n \mid \Delta\vec{d}_n, \Delta\vec{s}_n, \Delta\theta_n\right) \cdot p\left(\Delta\vec{d}_n, \Delta\vec{s}_n, \Delta\theta_n\right), \tag{32}$$

where Δc_n (temporal variation of the n th atom scalar product with the residual) depends on the choice of the new γ parameters. Considering $\Delta\vec{d}$, $\Delta\vec{s}$, $\Delta\theta$ independent, Eq. (32) turns into:

$$p\left(\Delta\gamma_n, \Delta c_n\right) = p\left(\Delta c_n \mid \Delta\vec{d}_n, \Delta\vec{s}_n, \Delta\theta_n\right) \cdot p\left(\Delta\vec{d}_n\right) \cdot p\left(\Delta\vec{s}_n\right) \cdot p\left(\Delta\theta_n\right). \tag{33}$$

Each of the probability functions has the form of a MRF. That is, they may be modeled by a Gibbs distribution [30]:

$$p(x) = \frac{1}{Z_x} \exp\left(-\frac{E_x\left(x\right)}{T_x} \right), \tag{34}$$

where $E_x\left(x\right)$ is an energy function that characterizes the MRF and how neighboring variables are related, while T_x stands for its variance.

From Eqs. (28), (31), (33) and (34) the functional to be optimized can be expressed as:

$$\begin{aligned} \Delta\gamma_n = \arg\min_{\Delta\gamma_n} \Bigg\{ & \frac{1}{2} \log\left(\left\| \hat{\mathcal{R}}_{n+1}^{t+1} \right\|^2 \right) + \lambda_{\Delta c_n} E_{\Delta c_n}\left(\Delta c_n\right) + \\ & \lambda_{\Delta\vec{d}_n} E_{\Delta\vec{d}_n}\left(\Delta\vec{d}_n\right) + \lambda_{\Delta\vec{s}_n} E_{\Delta\vec{s}_n}\left(\Delta\vec{s}_n\right) + \\ & \lambda_{\Delta\theta_n} E_{\Delta\theta_n}\left(\Delta\theta_n\right) \Bigg\} \end{aligned} \tag{35}$$

where $\Delta\gamma_n = \{\Delta\vec{d}_n, \Delta\vec{s}_n, \Delta\theta_n\}$ and each λ_x is a function of the statistics parameter T_x in Eq. 34.

B. Regularity Models

The general regularized expression to solve at every greedy iteration (35), requires the definition and modeling of each regularizing term E_x. In the following, the definitions of the Gibbs distributions arising in the MAP estimation are described together with the parametric modeling of the MRF.

1) Coefficient Model: Temporal variations of coefficients Δc_n should be small in ideal tracking of a primitive. In any case, coefficients may not change sign. Changes to coefficients should be driven mainly by the change of scale of the approximating function. To induce its temporal regularity, a normalized quadratic distance between the coefficients at time t and $t+1$ is considered for $E_{\Delta c_n}(\Delta c_n)$:

$$E_{\Delta c_n}(\Delta c_n) = \left(\frac{c_n^{t+1} - c_n^t \cdot \sqrt{\Delta sx \Delta sy}}{c_n^t \cdot \sqrt{\Delta sx \Delta sy}} \right)^2, \tag{36}$$

where previous c_n^t are re-normalized with respect to the scale transformation. The whole difference is normalized in order to make $E_{\Delta c_n}(\Delta c_n)$ independent of n, i.e. the mismatch error is proportional to the magnitude of the coefficients, and the coefficient magnitude has, in the ideal case, an approximately exponential decay with n.

2) Geometric Models: Displacement, change of scale and rotation constraints, are measured as the euclidean distance between the value under test and the most likely (ML) transformation estimated from previous MP iterations at every image location. Hence they can be represented as:

$$\begin{aligned} E_{\Delta \vec{d}_n} &= \left(d_x^n - \hat{d_x}^n \right)^2 + \left(d_y^n - \hat{d_y}^n \right)^2 \\ E_{\Delta \vec{s}_n} &= \left(s_x^n - \hat{s_x}^n \right)^2 + \left(s_y^n - \hat{s_y}^n \right)^2 \\ E_{\Delta \theta_n} &= \left(\theta^n - \hat{\theta}^n \right)^2, \end{aligned} \tag{37}$$

where $\hat{d}$, $\hat{s}$ and $\hat{\theta}$ correspond to the ML estimates (see. Sec. VI-D for details on their calculation). The use of motion information from the first appearing atoms to regularize the selection criteria of new ones, can be seen as a way to propagate the motion information from more reliable atoms to less reliable ones.

C. Setting of the Motion Model

Eqs. (36) and (37) define the potential among variables of the functional to optimize. However the model lays on the values assigned to the λ_x of Eq. (35). These values are unknown a priori and depend on the data to be analyzed since they represent the statistics that characterize the random variables Δc_n, $\Delta \vec{d}_n$, $\Delta \vec{s}_n$, $\Delta \theta_n$. In this work they have been considered to be constant for the whole sequence. Their value, as defined by Eqs. (34)-(37) is proportional to the standard deviation of the variables implied in the energy functionals described before. Hence, for a general sequence, their value needs to be trained. However, this will just be valid in average for the real transformation given the heterogeneous nature of motion in a general sequence. A detailed analysis of the proper adaption of a statistical model is out of the scope of this work. We focus on the understanding of the use of greedy approaches and parametric over-complete dictionaries for the approximation of image sequences.

D. Motion and Probability Fields Estimation

The transformation estimates are computed from all the atoms that interact in a certain region. In the example presented in this work (Eq. (8)) atoms have a localized support in space. Even though it is not strictly finite (see Fig. 7), amplitude decay is fast enough such that atoms located sufficiently far away can be considered as not interacting. Furthermore, the decay of the Gaussian envelop of (8) can be considered as well as an indicator that the strength of constraints (36) and (37) has to increase the closer an atom is from another, i.e. it is logical to consider that such 2 atoms must have a more coherent motion.

1) λ_x Modeling: From Eq. (8), the atom envelop is a bi-variate Gaussian with the same dimensions (sx, sy) as the atom itself:

$$\begin{aligned} p_\gamma(u,v) &= K \exp\left(-\left(u^2 + v^2 \right) \right) \quad \text{s.t.} \\ u &= \frac{\cos\theta\,(x-dx) + \sin\theta\,(y-dy)}{sx} \\ v &= \frac{-\sin\theta\,(x-dx) + \cos\theta\,(y-dy)}{sy}, \end{aligned} \tag{38}$$

where K is a constant. This model is assumed to represent also the influence law of the transformation of a given atom in a neighborhood. Thus, $\forall x, \lambda_x$ is proportional to $p_\gamma(u,v)$. The variance of the probabilities described in Sec. VI-A decreases depending on their relative distance and atoms scale. Thus, the model defined by the lambdas in section VI-C needs to have a local influence closely related to the structure of the signal and the set of functions used in each case to approximate a given frame. Theses λ_x values that model tightness in atom interactions are extended by their product with the bivariate model of (38).

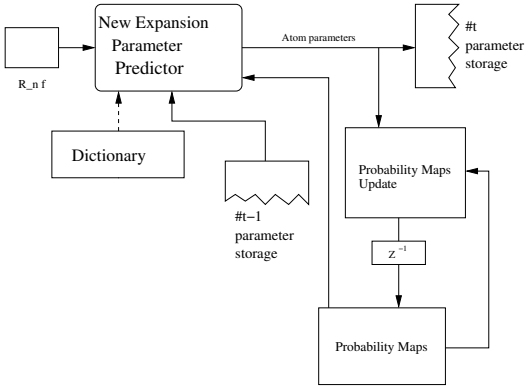

Fig. 6. Expansion Block Scheme.

2) Motion Parameter Estimates: The motion parameter estimates $\hat{d_x}, \hat{d_y}, \hat{s_x}, \hat{s_y}, \hat{\theta}$ of Eq. (37) are estimated from the preceding $n-1$ atoms of a frame expansion. They are the maximum likelihood estimates according to the energy probability associated to each atom.

In fact, considering that a given frame energy can be represented as the sum of the square of the coefficients in a MP expansion:

$$\|I_{t+1}\|^2 = \sum_{n=0}^{\infty} |c_n|^2 . \tag{39}$$

we may approximate the probability associated with the nth atom as the fraction of $\|I\|^2$

$$p(\gamma_n) = \frac{|c_n|^2}{\|I\|^2}. \tag{40}$$

The conditioned probability that a given atom contributes to spatial location (x, y) can be modeled through Eq. (37). Thus,

$$p(x, y \mid \gamma_n) = \frac{K}{\sqrt{sx \cdot sy}} \exp\left(-\left(u(x,y)^2 + v(x,y)^2\right)\right). \tag{41}$$

Hence, the motion parameters induced by atom $g^t_{\gamma_n}$ at point (x, y) have probability:

$$p(\gamma_n \mid x, y) = \frac{p(x, y \mid \gamma_n)p(\gamma_n)}{\sum_n p(x, y \mid \gamma_n)p(\gamma_n)}. \tag{42}$$

For localized atoms in space (as in the dictionary from sec. III-A), we can see that the summation in the above equation will only integrate those atoms close to position (x, y) due to their amplitude decay (Eq. (38)). Giving as example the case of the most likely displacement, or translational motion $E\left\{\vec{d} \mid x, y\right\}$ at a given (x, y), we formulate it as the average of all the transformations induced by all the atoms at a given point:

$$\hat{\vec{d}} = E\left\{\vec{d} \mid x, y\right\} = \sum_n \hat{\vec{d}}(x, y)_n \cdot p(\gamma_n \mid x, y). \tag{43}$$

The same applies to the remaining geometrical parameters $\hat{\vec{s}}, \hat{\theta}$.

In Eqs. (39) - (42) the whole set of terms for the expansion of I are considered. However in a practical and realistic situation, only a truncated version of the representation can be considered. Indeed, to estimate (predict) the motion that the nth atom will follow, only the precedent $(n-1)$ available atoms are considered for the statistical measurements and calculations.

VII. Rate-Distortion Formulation

A similar framework to Sec. VI can be considered in terms of a rate-distortion (R-D) functional. As in Sec. VI regularity assumptions can be taken into account to exploit the redundancy in the representation of the geometrical transformation of neighboring atoms. In this case regularity would be somehow measured by the rate. The optimization to be solved corresponds to jointly minimizing distortion and rate:

$$\min_{F_N^t} \{D_N + \lambda R_N\}, \tag{44}$$

where N represents the number of terms in the expansion to represent a given frame. From the properties of matching pursuit representations (sec. III-A),

$$\begin{aligned} D_N &\leq \sum_{n=0}^{N-1} |\xi_n|^2 + \left\|\mathcal{R}^N I_{t+1}\right\|^2 = \\ &\sum_{n=0}^{N-1} |\xi_n|^2 + \|I_{t+1}\|^2 - \sum_{n=0}^{N-1} |c_n|^2 = \\ &= \|I_{t+1}\|^2 - \sum_{n=0}^{N-1} \Delta D_n, \end{aligned} \tag{45}$$

where I_{t+1} is the original frame to be approximated and ΔD_n corresponds to the contribution to reduce the distortion of atom n, thus $\Delta D_n = D_n - D_{n-1}$. In Eq. (45) ξ_n corresponds to the quantization error of the coefficients c_n, and can be assumed to be independent of the coefficient.

If r_n is taken as the total cost needed to code the nth term of the expansion, then it follows from Eqs. (44) and (45), that Eq. (44) can be upperbounded as

$$\min_{F_N^t} \{D_N + \lambda R_N\} \leq \min_{F_N^t} \left\{ \|I_{t+1}\|^2 - \sum_{n=0}^{N-1} \Delta D_n + \lambda \sum_{n=0}^{N-1} r_n \right\} = \min_{F_N^t} \left\{ \hat{D}_N + \lambda R_N \right\}. \tag{46}$$

Considering $J_N(\lambda) = \hat{D}_N + \lambda R_N$ then

$$\begin{aligned} &\min_{F_N^t} \{J_{N-1}(\lambda) - \Delta D_N + \lambda r_n\} = \\ &\min_{F_N^t} \{J_{N-1}(\lambda) + \Delta J_N(\lambda)\} = \\ &\|I_{t+1}\|^2 + \min_{F_N^t} \left\{ \sum_{n=0}^{N-1} \Delta J_n(\lambda) \right\}. \end{aligned} \tag{47}$$

Thus, a compact representation of the problem is:

$$\min_{F_N^t} \left\{ \sum_{n=0}^{N-1} \Delta J_n(\lambda) \right\}. \tag{48}$$

Such a formulation implies a global optimization which, depending on the dictionary (e.g. non-orthogonal) and optimization constraints (e.g. non-divisibility in orthogonal smaller sub-problems), may be of overwhelming complexity (see Sec. III-B). In the scope of MP, Eq. (48) turns into a suboptimal solution where every ΔJ_n is minimized at every iteration. This can be considered as the criteria for selecting $g_{\gamma_n}^{t+1}$:

$$\min_{F_N^t} \left\{ \sum_{n=0}^{N-1} \Delta J_n(\lambda) \right\} \leq \sum_{n=0}^{N-1} \min_{F_{\gamma_n}^t} \{\Delta J_n(\lambda)\}. \tag{49}$$

Indeed, for general redundant dictionaries, $\Delta J_n(\lambda)$ $n \in N$ are not necessarily independent among them. *Distortion* reduction and *Rate* investment at a given state of the WMP algorithm may depend on previous iterations. Closeness to optimality will be conditioned by the structure of the dictionary and how this relates to the signal to approximate.

VIII. Implementation Issues

A. Low Frequency Representation

The dictionary presented in Sec. III-A.2 has been designed for efficently representing edges. We thus use it in order to approximate the high frequencies of the signal after having removed the low frequency approximation by means of a Laplacian pyramid.

B. Atom Refresh

All the information appearing in a frame at time t can not be mapped from the previous frame and vice-versa. Indeed, we consider a forward mapping scheme where all atoms from frame at time $t-1$ try to get matched in the frame at time t. This is not always possible and sometimes the atom will not be able to find at t the feature it was representing at time $t-1$. In the present approach we consider a measure of the reliability of the prediction of a given atom evolution. At every new frame the normalized scalar product of the transformed atom is compared with the initial projection of the atom within the first frame.

$$\left| \frac{\left\|c_n^{t+1}\right\|^2}{sx_n^{t+1} sy_n^{t+1}} \right| \geq \frac{\|c_n^0\|^2}{sx_n^0 sy_n^0} \cdot \delta, \qquad \delta \in (0,1] \tag{50}$$

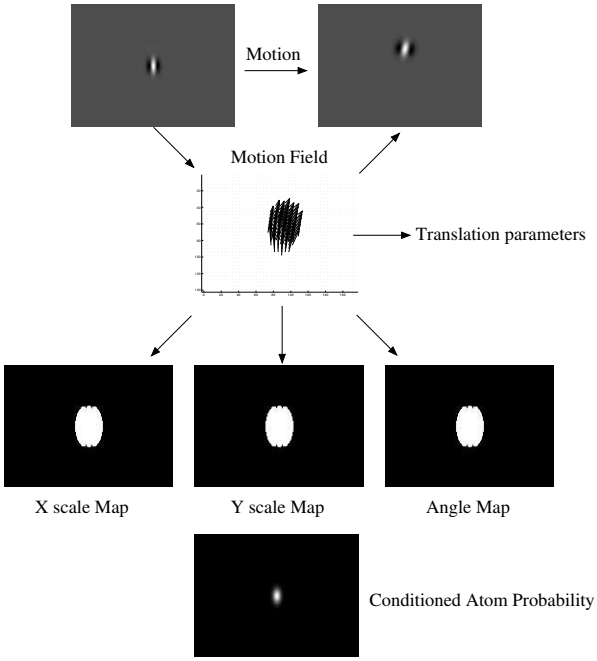

Fig. 7. Atom transformation maps and parameters. The parameter maps (X scale, Y scale and Angle) correspond to the areas where the geometry of a selected atom will influence future greedy iterations.

If a significant drop in the scalar product is detected the atom is canceled (the trajectory is not valid anymore). At the end of the projection process, those atoms that have been canceled are reintroduced in the frame through a full MP search. In the investigation performed in this work, the atom refresh has been set such that a very small portion of atoms can be renewed at every frame (e.g. no more than three percent).

C. Motion Initialization

The functions in use for the generation of our dictionary have a relatively simple shape. In the direction parallel to contour gradients, very likely represented by the smooth part (Gaussian) of Eq. (8), even very relevant atoms may slide: this is similar to the well know "aperture" problem. To avoid this, a first initialization of the expected motion maps is essential. Thus, in the case of no *a priori* indicator of the motion of a primitive, the whole pixmap of the original image included in the support of that primitive is used for a first estimate. The cross-correlation (matching) of the zero mean and normalized versions of the patch and the frame that we want to approximate is used, i.e. the correlation between the normalized patch and the normalized frame is measured for every possible geometric transformation of the atom.

D. Motion Maps Update

A set of geometrical parameter maps are kept during the iterative decomposition of a frame. These contain the local geometrical deformations that atoms suffer in their adaption to represent a new frame. Geometry maps are updated progressively at each iteration of the greedy algorithm. After the retrieval of a function, its transformation parameters are introduced in the maps as described in VI-D.2. The information of the maps is used to introduce regularity in the selection procedure at every greedy iteration. In this way, the motion registered by the first atoms found in a certain image area will condition the geometrical deformation of posterior atoms found in that area. In Fig. 7 a representation of an atom transformation can be seen together with the associated motion. We show as well the influence area where the parameter maps will be considered. At the bottom, we show the conditioned probability (an-isotropic Gaussian) that will take part in the computation of the most likely local motion transformation given an image location.

IX. Experimental Results

In this section we present several results corresponding to the effect of regularization on different sets of sequences, both synthetic and natural. The results shown in the way of vector fields correspond not only to the translation of atoms but also to their deformation, i.e. the interpolated motion of atoms is represented on their whole support (see Fig 7). These representations (Figs. 9,8,11,13), although they may suggest an optical flow meaning, must be interpreted more in the sense of atoms flow. Some examples of transformation are given for particular atoms (Figs. 9 and 13). In order to illustrate in a more objective way the effects of regularization, in IX-C a simple coding scheme is used to code the parametric sequence representation and some R-D results are given for the foreman sequence in QCIF format.

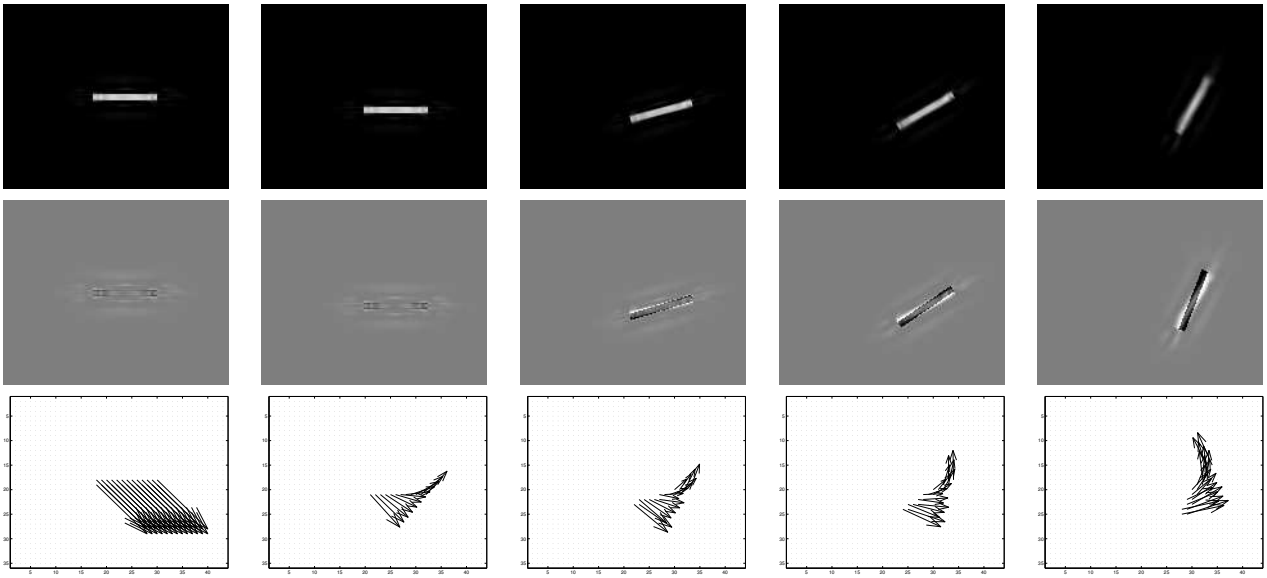

Fig. 8. Affine motion of a synthetic model (line). From top to bottom: approximation of the line, residual with respect to the original model and motion associated to the atoms.

A. Synthetic Sequence Examples

In Fig. 8 we show an illustration of the proposed paradigm based on steering image primitives through a sequence. In this test, like in all the rest, the dictionary in use is the proposed one in Sec. III-A. The flow represented in the third row shows how atoms transform to follow and match the successive transformations of the sequence. Just above in the figure, the resulting approximations and the residuals show that although primitives adapt better to shape and motion trajectory, they are subject to the lack of resolution of the dictionary. The effects can be seen in the evolution of the residual error after the approximation. In fact, in the absence of regularity constraints, atoms try to reorganize themselves in order to reduce this residual error. Consequently, this would force many terms of the successive frames expansions to reorganize in position, angle and scale, leading to a irregular representation of the motion transformation. As part of the approximation error, a progressive blurring of the edges of the rod can be observed. This corresponds to the adaption of function coefficients. They try to reduce the negative effect of slight motion parameters mismatch, together with the sub-optimality of the greedy approach. Hence, a cumulation of prediction error appears through time.

Continuing with another synthetic sequence, we can see this time in Fig. 9 the example corresponding to the motion associated to a particular atom. The sequence corresponds to a translating and rotating square. We consider a certain atom, represented in the picture by a white mark that has the shape of its support. In both columns, we see the representation of the square by means of a 50 coefficients expansion with the footprint of the function support superimposed. In the left column we display the corresponding past and present positions of the atom for the non regularized case, i.e. the selected atom is fully driven by the search of the highest projection coefficient absolute value. On the right, the atom is steered considering the a priori of rigid motion. At the bottom of Fig. 9 we can see the motion associated to atoms of the right column.

B. Natural Scene Examples

The synthetic models considered in the figures above are very simple and constrained. Fig. 10 shows a comparison between a non regularized result (left), and a regularized one (right). A clear influence of the regularization and motion initialization is reflected in the flow related to the atoms motion. Figures on the right show a clear relation can be established between atoms that participate in the cars approximation and their motion. In the example where the truck appears, the influence between neighboring atoms located in the wood area in the background can not be avoided, i.e. the moving atoms of the truck *push* in some measure the atoms representing the background. Interdependence among neighboring primitives is responsible for their strong interaction.

In Fig. 11, a set of consecutive approximated frames appears together with the atoms motion flow. Notice the regularized motion of atoms that follows the object trajectories. However, interactions among neighboring atoms are observed, uncovered and covered parts in some situations may enter in interaction. Indeed shrinking, dilation or slight displacement can manifest due to MP sub-optimality, dictionary granularity and the interaction among atoms. The

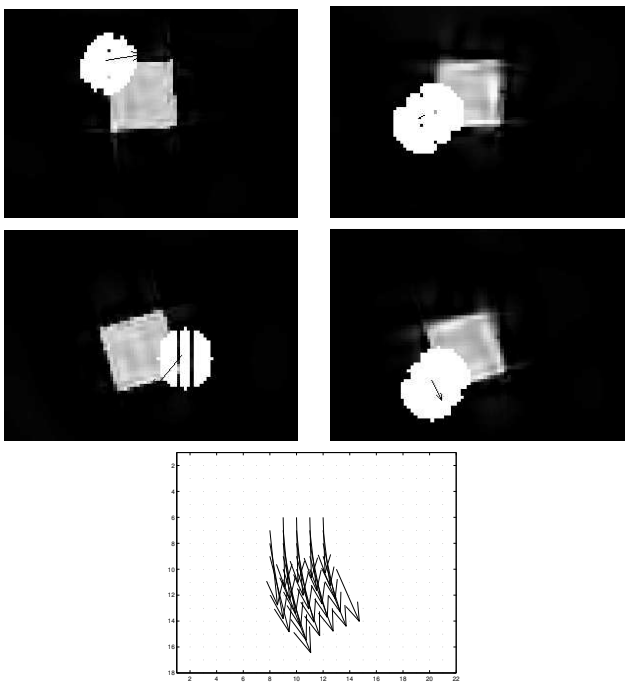

Fig. 9. Affine motion of a synthetic model (square). The white batch corresponds to the footprint of a selected atom in two temporal instants. Left is the non-regularized prediction. Right is the regularized prediction. Bottom: most reliable motion of the regularized solution (atoms flow in the area where atoms amplitude is significant). Rotation and displacement can be appreciated.

WGA acts as if the dictionary in use was modified to approximate the signal, i.e. the approximation fidelity of the signal is conditioned by the constraints imposed in the atom search, dictionary parameters sampling and MP sub-optimality. This is reflected in Fig. 12 where a comparison with the non-constrained greedy local search is performed. A price is paid for the regularization of atom transformations, a cumulative loss is appreciated due to the successive mismatch. Indeed the effect of the WGA is to trade between regularity and signal approximation.

In order to illustrate the behavior of atoms in a natural sequence, we show in Fig. 13 a set of pictures that represent from top to down: the original set of images, the approximation using 500 atoms, the flow of atoms, and the evolution of 3 different atoms trough time. The atoms flow shows the motion of image primitives, as well as their deformation. Changes suffered by those representing the background to adapt to the motion of the head are clearly appreciated. This is because several big atoms are used to describe the background. Certainly, the building appearing behind the head can be represented very efficiently with long oriented atoms with a large scale in the parallel direction to the lines. These have to re-adapt their parameters to fit as well the motion of the head. The three atom examples appearing in the last three rows are composed by the approximation pictures plus the footprint of the represented atom. They capture the motion of the head. Translation, rotation and scaling can be appreciated on them (Fig. 13). We must notice that atoms reintroduced by means of a refresh are not taken into account in the flow representation.

Evidence of the regularization effects in the sequence *foreman* can be found in Fig. 14. Atoms deformation becomes less instable with the regularization. Notice how important this effect is on the region where the detail of the lines of the building are. In the absence of regularization and motion initialization atoms motion is not accurate in their smooth direction. Furthermore, MP facilitates the propagation of error to neighboring atoms in the area. Motion initialization by means of matching is a key element for stability. However, this comes at the price of a progressive temporal drift in the prediction of frames (Fig. 15).

C. Effect of Regularity in a Coding Sense

In order to have an objective measure of the regularization effects, we consider the R-D curve obtained for a simple coding scheme applied to the parametric representation of the foreman sequence. Every frame is described by a set of atoms which are obtained sequentially by the iterative greedy algorithm. The criteria used in the function selection rule of our algorithm is designed such that a spatial and temporal regularity is imposed among atoms. Hence, correlation of atoms at time t with their evolved version at time $t+1$ will be exploited by only encoding the parameters and coefficients differences. When an atom is refreshed, this is obtained by doing a full search in the whole image (as

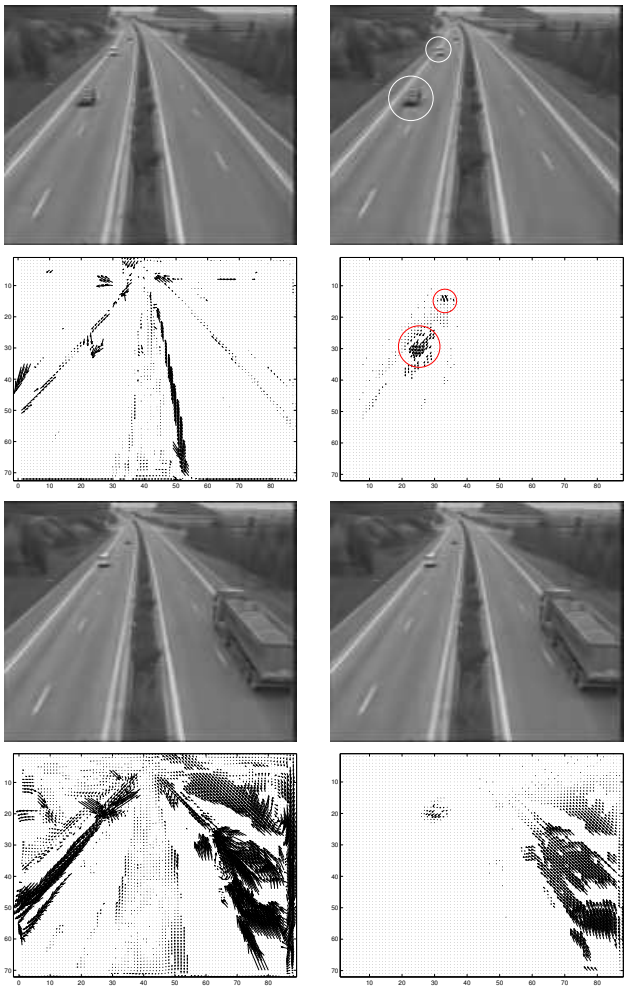

Fig. 10. Natural sequence motorway. Left column: non-regularized solution. Right column: regularized tracking. First and third rows: Respective reconstructions with 500 atoms. Second and fourth rows: Most reliable primitives motion

described in sec. VIII). Atoms that have been refreshed will also be coded by just sending the difference with respect to the atom they replace in the previous frame. Finally, an arithmetic coding of the differential data is performed.

The curves on Fig. 16 show the gain obtained in terms of R-D of the regularized Bayesian matching with respect to the non-regularized one. The use of regularization in the matching criteria imposes a certain structure among the behavior of atoms in a frame. This helps reducing the instability of image primitives. A consequence of the regularization turns to be, as expected, the reduction of the amount of necessary bit-rate to represent frame to frame variations. The entropy of the parametric representation gets reduced by the low-pass filtering of parameters imposed by the MRFs criteria. Furthermore, MRFs criteria and the motion initialization when no a priori is available reduces the propagation of error in the atoms parameters, contributing to a better D-R behavior. However, and as shown in Fig. 15 this is in exchange of a higher drift. A 3dBs loss is cumulated through the GOP. This cumulates due to the limited dictionary resolution, the sub-optimality of MP and the inherent prediction effects. The variation of rate presented in the curves is obtained by exploiting the natural SNR scalability that MP expansions have. For a given bit-rate, video frames are progressively reconstructed by limiting the number of atoms used per frame such that the coding cost respects the selected bit-rate. Thus, with less regularity, coding costs are higher and a reduced number of atoms can be considered for decoding.

Figure 17 shows the effect of the regularization in terms of rate distortion for the foreman sequence. Both curves show the common drift behavior that comes up from the predictive nature of the representation. Notice that the regularized version has a gain between 0.5-1.5 dBs over the non-regularized with 20kbps less. Notice the difference between Fig. 15 and Fig. 17. The weakened matching criteria of the greedy algorithm produces a loss in the regularized approximation with respect to the non-regularized one when the same number of atoms per frame is used in both cases (Fig. 15). That is because, in the regularized case, atoms are not able to freely move and place themselves to compensate errors

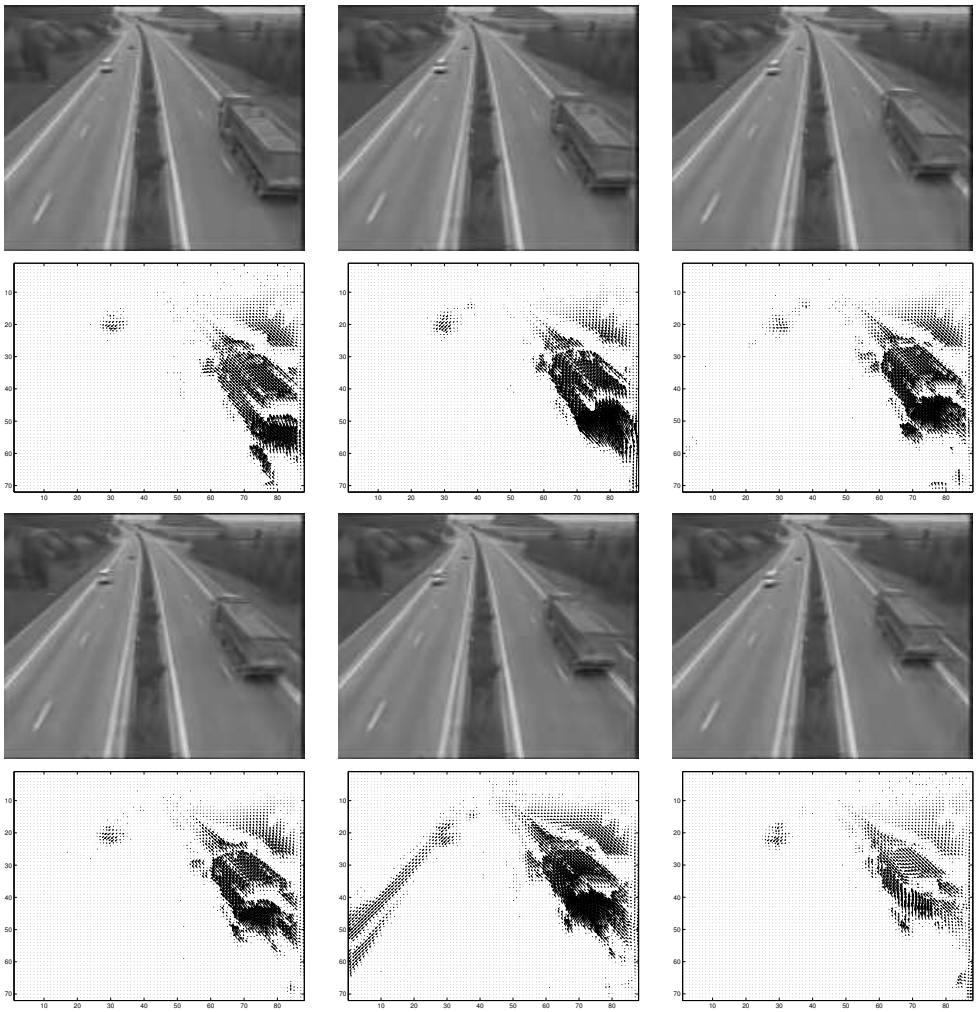

Fig. 11. Several consecutive frames of a natural sequence showing the reconstructed signal with 500 coefficients with the functions associated motion. The transformation of atoms from frame to frame was done using the criteria with a priori information.

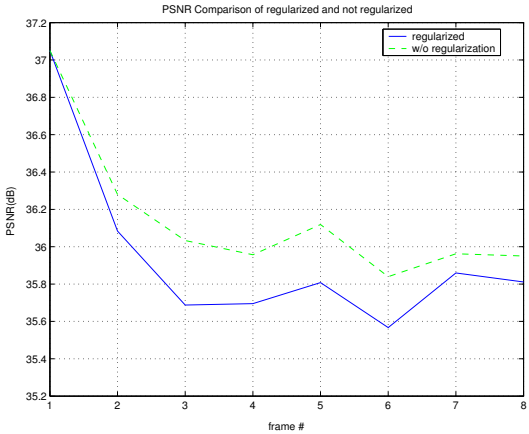

Fig. 12. Curves representing the loss from frame to frame (corresponding to those of Fig. 11) approximation accuracy due to the regularization of the function parameters.

done by earlier atoms of the iterative decomposition. Indeed, parameters quantization introduce motion missmatch in the atoms deformation. However, the reduction in the variability of the deformation parametrization turns into a reduction of the coding costs allowing to decode more atoms per frame from the regularized stream (Fig. 17).

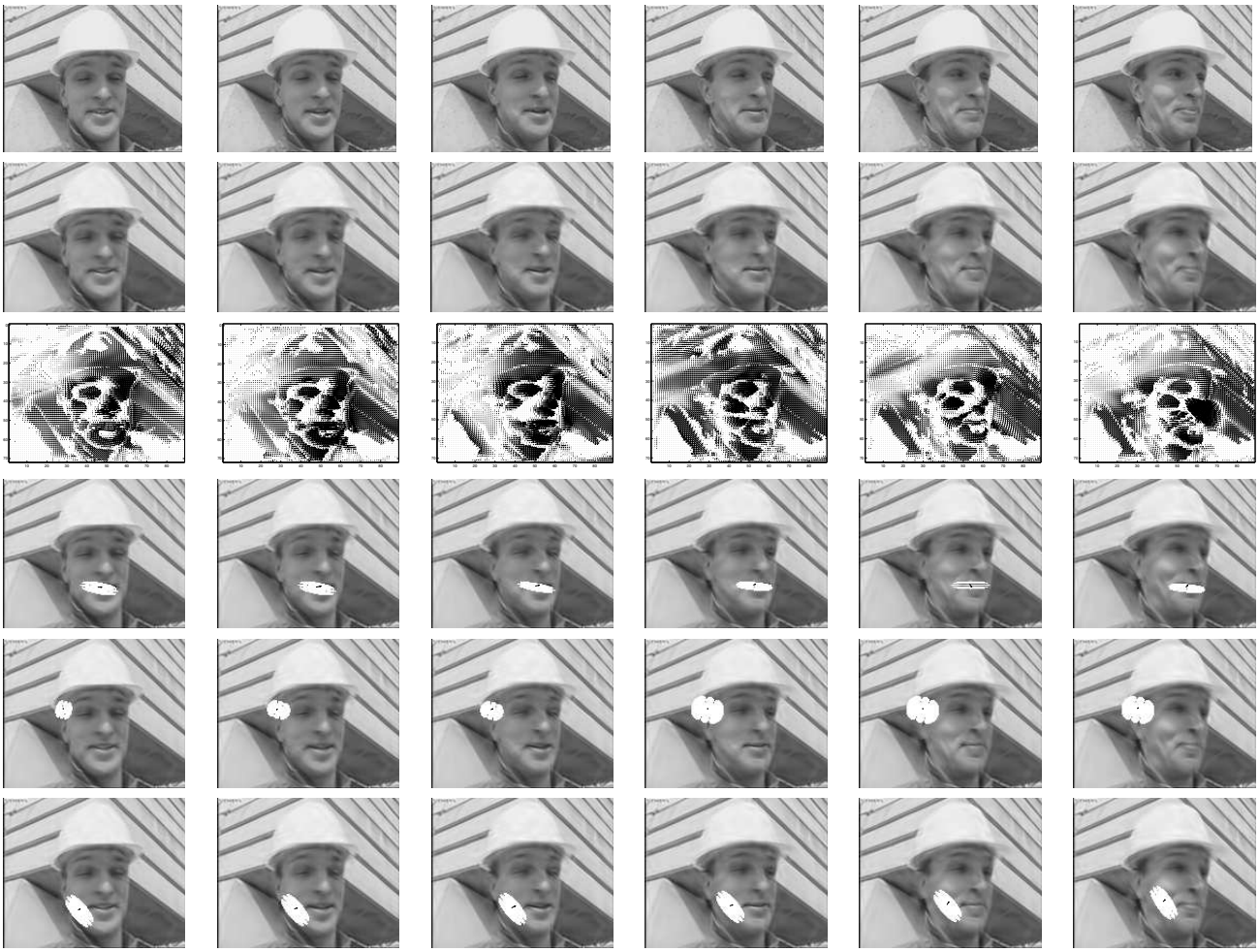

Fig. 13. Several consecutive frames of a natural sequence showing the reconstructed signal with 500 coefficients with their associated motion. It can be seen from top to bottom, the original frame, the reconstructed approximations, the motions associated to the functions,and the motion of 3 different atoms in the sequence and the region where they contribute.

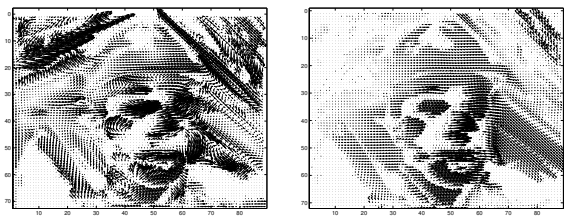

Fig. 14. Comparison of the computed deformations (atoms associated motion) for the 2nd frame of the foreman sequence: left not regularized, right regularized.

Temporally consecutive atoms that are sequentially predicted do not necessarily last all along a GOP. Fig. 18 shows the histogram of temporal lengths for atoms prediction that are determined by the criteria described in sec. VIII. In this example the 48 first frames (3 GOPS) of the sequence foreman have been taken into account for the generation of the statistics. The total number of temporal atoms (sets of atoms that are predicted from frame to frame without being refreshed) in this 3 GOPs is 1876. There are about 35 per cent of atoms that succeed in being predicted from frame to frame during all the GOP. However, a relevant number need to be refreshed quite often, common temporal lengths are from 1 to 8 frames. Sequence changes (occlusions, uncoverings and simple interaction among atoms) force their refresh. Atom refresh is a natural manner to introduce components to represent new information that appeared in the signal. However, atom interaction as well as mismatch due to the lack of resolution on the function parameters, contributes to the unnecessary rising of the refresh rate. Hence coding efficiency is reduced.

Atom refresh rate is not the same for every temporal frame, depending on the motion of the picture and the drift propagation through MP iterations, a different number of coefficient primitives pass the threshold fixed to determine whether an atom can still be considered worth to be kept or not. In Fig. 19, we monitor the number of refresh atoms

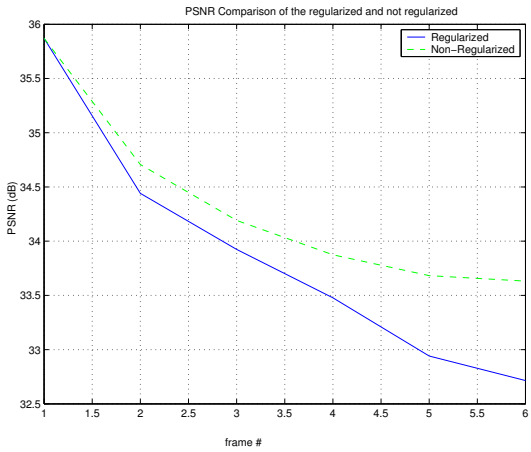

Fig. 15. Curves representing the loss of frame (from Fig. 13) approximation accuracy due to the regularization of the function parameters.

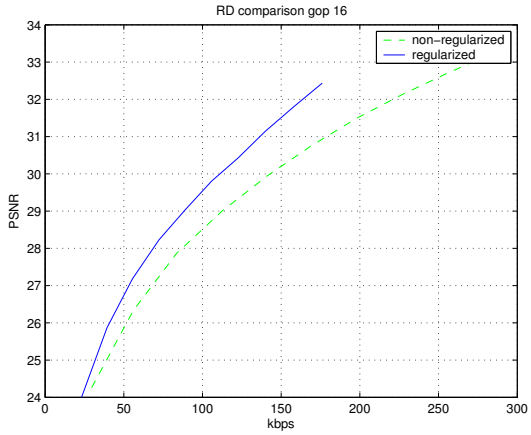

Fig. 16. Comparison of the regularized and non-regularized foreman sequences (16 frames GOP).

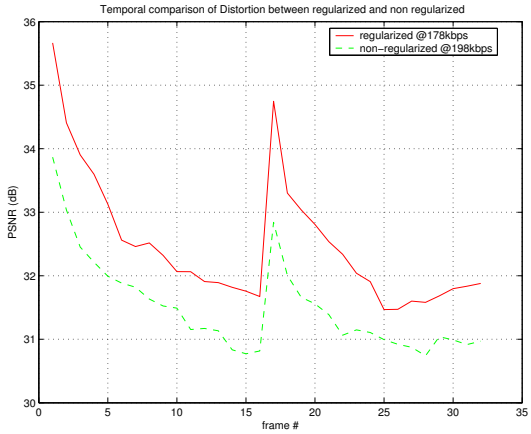

Fig. 17. Comparison for the regularized and non-regularized cases of the foreman sequence with GOP 16.

introduced at every frame. As shown, intra frames (the first at every GOP) are not considered in this graph. In the graph a zero every 16 frames appears even if all atoms are *refreshed*, i.e. non of them is predicted from the previous frame. In Fig. 19 several maximal pics more relevant than others can be identified. Close relations can be found between these and sets of frames of the sequence where most relevant changes appear. In the first GOP, the head appearing in the sequence turns from left to right. This movement progressively occludes a part of the face. At frame 12, the face starts to turn back to its initial position, uncovering in the procedure areas that were not visible before. This requires the insertion of additional information to cope with the change of topology of the contours and shape in the picture,

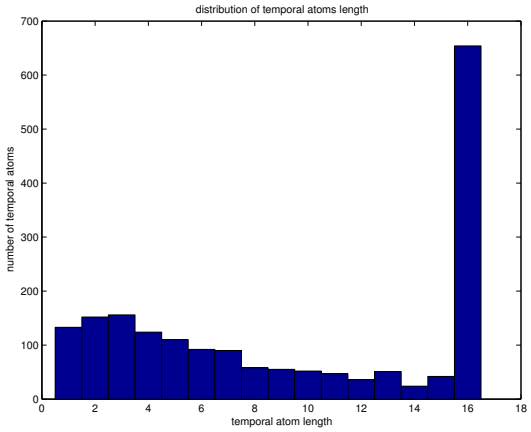

Fig. 18. Distribution of length for the temporal atoms. The length is determined by the atom refresh criteria of sec. VIII where atoms loosing 80% of their amplitude are refreshed

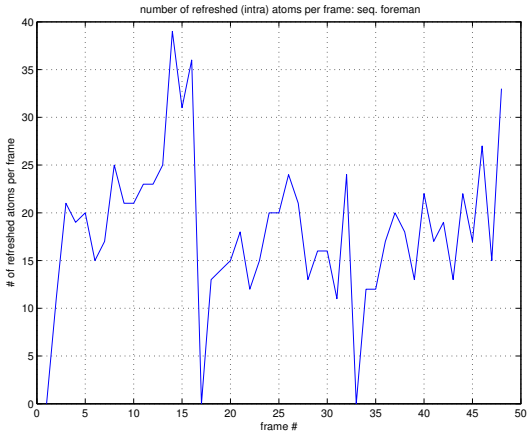

Fig. 19. Number of new atoms introduced in the refresh procedure in each frame of the sequence.

i.e. deformation of atoms located nearby the changing area in frame 12 can not represent properly the new appearing regions in frame 13 (Fig. 20).

X. Discussion

At first spatio-temporal representation of video sequences based on evolving geometry primitives could be approached as in the case of images (sec. III-A). A 3D dictionary g_γ^{3D} could be defined where functions, in addition to geometrical spatial information, would also model the temporal evolution. Such a dense dictionary could be used in a greedy decomposition approach as a generalization of [31]. However, the temporal sampling of video signal is so low that simple models like the linear (e.g. constant speed displacement) are unlikely to adapt efficiently to the somewhat complex trajectories. Furthermore, the complexity of considering all the possible affine geometrical transformations that a given primitive may suffer through time rapidly becomes highly prohibitive, i.e. the whole dictionary needs to be browsed at each iteration. At the same time, data volume is much bigger than in the 2D case. A part from

Fig. 20. Reconstructed frames 12,13 from the foreman sequence. In them we observe the uncovering of the left region (right in the picture) of the man's face.

that, in some applications, 3D approaches may be inappropriate due to delay issues. This motivates the present study where an approach based on the progressive projection of primitives from frame to frame is carried. The goal is to get less data involved for the signal expansion. Hence, a predictive approach seems to be indicated where a frame to frame adaptation of the image geometrical features is performed. The direct consequence of this approach is that a general consideration of the whole set of frames is not available anymore. This makes it unlikely to retrieve an optimal solution for the whole GOP at a given MP iteration. Primitives are selected in the basis of a single frame and then projected to others hoping to find the temporal evolution of these. As presented in this work, very simple primitives from a very redundant dictionary and the use of MP require the knowledge of a reliable a priori guess, e.g. by means of pixmap matching. Thanks to their geometrical meaning, primitives are able to catch the essence of the structure of images. However, two main factors spoil the prediction of frames by the motion or deformation of primitives. These have to be taken into account in order to understand which limitations affect the present approach and what has to be handled in order to profit from the flexibility of MP schemes while reducing their drawbacks. The first main issue comes from the assumption of "uniform" motion in the area of influence of a function and the second is the high coherence between some neighboring functions. When the assumption of "uniform" motion fails, a mismatch appears between the deformed function and the transformed image. This mismatch introduces a drift in the residual image used for the next MP iteration. The consequence is a modified initial condition to find the next atom in the expansion procedure. Thus, an atom different than the one required by local motion may be selected in order to compensate the drift. This influence propagates the mismatch to later iterations, due to coherence among functions. In the case where the best motion for an atom is known in advance (e.g. estimated by another method and introduced in the atom selection by means of a MAP decision criteria) and this motion considers the atom to be independent from others, the assumption of motion consistency on the support of a function may nevertheless fail at some point. Consider the following example: We have discussed before that BM can be seen as a particular case of the approach considered in this work. Every block of the predicted frame is approximated using a given pixmap block belonging to a set of possible blocks from the previous frame. In this particular case, a certain mismatch on the prediction of a block will not influence the rest since orthogonality is normally imposed among predicted blocks. Let's consider now another example where translational motion obtained from a BM scheme is applied to drive the atoms in a MP approach. BM motion would be the a priori information for the displacement of atoms from frame to frame in a sequence. Atoms with a large scale having a significant interaction with atoms placed in neighboring blocks can be the cause of mismatch when neighboring motion estimation blocks have different motion directions. Under different displacement, two atoms that where complementary will lose their alignment. Unlike in adaptive block matching motion compensation, here we do not consider the division of a function into smaller ones to better locally adapt to motion when the latter is not uniform in all the matching area.

XI. Conclusions

The present work explores the possibility of representing video sequences by the parametric deformation of 2D geometrical primitives to approximate each frame. A generic formulation of the problem has been investigated showing its relation with known non-linear signal approximation approaches. The problem of sequence representation has been formulated as the minimization of the approximation error based on a inverse problem subject to a constraint that depends on the application (regularity, rate, etc...). Furthermore, an approximate solution can be obtained by means of a constrained greedy algorithm. Theory and experimental results justify the need for a priori knowledge to help the decision criteria of the greedy approximation when general over-complete dictionaries are in use. The main issues to face are the simplicity of waveforms used in the parametric motion prediction from frame to frame and the stability of the greedy approach when using over-complete dictionaries. This concerns the robustness and accuracy of the parametric representation estimation of frame to frame transformations. Present results with MRF suggest that in order to use dictionary functions to retrieve the transformations from frame to frame, the dictionary should be adaptively defined at every frame prediction by grouping appropriately atoms together. This should exploit the particular shape (or geometrical qualities) of a region since more complicate structures would be implied in the matching procedure. Hence, contributing to a higher discrimination among possible motions. Sampling in the space of parameters involved in the definition of the dictionary is also of capital importance in the approximation of the geometrical transformations of a group of functions. In fact, the selection of a parameter sampling resolution will trade between representation accuracy, complexity and, in the case of a coding application, rate. In the present work we have considered for simplicity (and important complexity issues) a scheme based on a forward prediction. In order to suitably stop tracking functions that are occluded, and consider new appearing information a backward strategy would be preferred. This should be feasible (in terms of complexity) in a framework using sets of atoms that move in a rigid manner, in such case, the number of atoms sets should be much smaller than the number of functions used to describe a frame. This can make the backward scheme affordable and a global optimization possible, i.e. a solution by means of MP would still suffer of the stability problem for coherent sets of atoms. In the absence of a globally optimal solution, additional reliable a priori motion information would still be necessary.

Appendix

A. Block Dictionary MP Stability

Consider the situation posed in Sec. V-B where the dictionary $\mathcal{D}$ is the union of several sub-dictionaries such that:

$$\mathcal{D} = \bigcup_{i=0}^{N-1} D_{B_i}. \tag{51}$$

Let f be a function such that

$$f \in span\left(g_{\gamma_{B_i}} : i \in [0, m-1],\ g_{\gamma_{B_i}} \in \mathcal{D}_{B_i}\right), \tag{52}$$

i.e. f can be expressed as a linear combination of atoms $g_{\gamma_{B_i}}$ where no more than one primitive is taken from each dictionary block $\mathcal{D}_{B_i}$. This is a very restrictive situation. However several examples can be found in practice as depicted in Sec. V-B where this situation may apply. Given the additional constraints imposed to the dictionary and the signal f, a refinement of the exact recovery condition (Stability Condition) defined in Theorem 1 can be established.

A new measure on the coherence can thus be introduced where the block based division of the dictionary is taken into account. Borrowing ideas from [25] where the same coherence measure was used for a similar situation, we define the *Babel block* function $\mu_{1_B}(m)$.

Definition 3: Let $\mathcal{D} = \bigcup_{i=0}^{N-1} D_{B_i}$ denote a block dictionary and Γ the set of sub-blocks from where, at most a function is taken from each, then the cumulative coherence function $\mu_{1_B}(m)$ is

$$\mu_{1_B}(m) \triangleq \max_{\Gamma|\ \|\Gamma\|_0=m} \max_{j\notin\Gamma, l} \sum_{i\in\Gamma} \max_{k} \left| < g_k^{B_i}, g_l^{B_j} > \right|, \tag{53}$$

This measure defines the worst cumulative dot product possible among two functions of different blocks for the worst selection of the Γ set of blocks.

We can now prove Theorem 2.

Proof:

From (20) and following the procedure suggested in [11], it follows:

$$\begin{aligned} \sup_{g_\gamma \notin D_\Gamma} \left\| (D_\Gamma)^+ g_\gamma \right\|_1 &= \sup_{g_\gamma \notin D_\Gamma} \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} D_\Gamma^T g_\gamma \right\|_1 \\ &\leq \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{1,1} \sup_{\gamma \notin D_\Gamma} \left\| D_\Gamma^T g_\gamma \right\|_1 \end{aligned} \tag{54}$$

The first term that corresponds to the $1,1-norm$ of the inverse of the Gram matrix can be expressed as:

$$\left(D_\Gamma^T D_\Gamma\right)^{-1} = (I + A)^{-1}, \tag{55}$$

where I denotes the identity matrix and A all the off-diagonal components of the dictionary Gram matrix. Expanding (55) by means of Von Neumann series and using $\|A\|_{1,1}<1$ we have:

$$\left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{1,1} = \left\| \sum_{k=0}^{\infty} (-A)^k \right\|_{1,1} \leq \sum_{k=0}^{\infty} \|A\|_{1,1}^k = \frac{1}{1 - \|A\|_{1,1}}. \tag{56}$$

$\|A\|_{1,1}$ is the biggest $\updownarrow_1$ norm column of matrix A and

$$\|A\|_{1,1} = \max_k \sum_{j\neq k} | < g_k, g_j > | \leq \mu_{1_B}(m-1). \tag{57}$$

The second term of (54) term can be upper bounded as follows:

$$\sup_{g_\gamma \notin D_\Gamma} \left\| D_\Gamma^T g_\gamma \right\|_1 = \sup_{g_\gamma \notin D_\Gamma} \sum_{g_l \in D_\Gamma} | < g_\gamma, g_l > | \leq \mu_{D_B} + \mu_{1_B}(m-1), \tag{58}$$

where μ_{D_B} denotes the maximum coherence (inner product) between two functions into a block. In order to be more explicit, (58) represents the worst possible case for the addition of cross products between the optimal set of m atoms and an atom not belonging to this set. This is, the $m-1$ worst possible cross products between atoms belonging to different blocks of $\mathcal{D}$ plus the worst possible coherence between two atoms belonging to the same block.

Hence, exact recovery of the the right set of atoms is ensured when:

$$\frac{\mu_{D_B} + \mu_{1_B}(m-1)}{1 - \mu_{1_B}(m-1)} < \alpha \tag{59}$$

■

B. Model Based MP Stability

Consider the existence of an *a priori* guess ($W(f)$) for the atom selection that depends on the function f. The inner product between the residual at a given MP iteration and the dictionary can be interpreted as the probability of a given function to be selected at a certain iteration

According to [11] we see then that given the sub-dictionary D_Γ where columns are the "correct" linearly independent functions of the sparse representation of f, at every iteration, the following should be satisfied for a $Weak(\alpha)$ greedy algorithm:

$$\rho\left(R_n\right) = \frac{\left\|W_{\overline{\Gamma}}\left(D_{\overline{\Gamma}}^T R_n\right)\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T R_n\right)\right\|_\infty} < \alpha, \tag{60}$$

where $D_\Gamma \subset \mathcal{D}$ and $D_\Gamma \cup D_{\overline{\Gamma}} = \mathcal{D}$ and W_Γ, $W_{\overline{\Gamma}}$ are two diagonal matrices containing the weights $w_{ii} \in [0,1]$ that define the *a priori* information about the suitability of that function in the sparse representation of f.

A new Stability Condition (SC) can thus be formulated as estated in Theorem 3.

Proof:

According to the assumption that $R_n \in \ span(D_\Gamma)$ and that columns of D_Γ are linearly independent, then $R_n = \left(D_\Gamma W_\Gamma\right)\left(D_\Gamma W_\Gamma\right)^+ R_n = P_\Gamma R_n$. This gives:

$$\begin{aligned}\frac{\left\|W_{\overline{\Gamma}}\left(D_{\overline{\Gamma}}^T R_n\right)\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T R_n\right)\right\|_\infty} = \frac{\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(D_\Gamma W_\Gamma\right)\left(D_\Gamma W_\Gamma\right)^+ R_n\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T R_n\right)\right\|_\infty} = \\ \frac{\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(\left(D_\Gamma W_\Gamma\right)^+\right)^T \left(D_\Gamma W_\Gamma\right)^T R_n\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T R_n\right)\right\|_\infty}.\end{aligned} \tag{61}$$

This quantity can be bounded by:

$$\begin{aligned}\frac{\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(\left(D_\Gamma W_\Gamma\right)^+\right)^T \left(W_\Gamma D_\Gamma^T\right) R_n\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma R_n\right)\right\|_\infty} \leq \\ \left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(\left(D_\Gamma W_\Gamma\right)^+\right)^T\right\|_{\infty,\infty} = \\ \left\|\left(D_\Gamma W_\Gamma\right)^+ \left(D_{\overline{\Gamma}} W_{\overline{\Gamma}}\right)\right\|_{1,1}.\end{aligned} \tag{62}$$

Given that $\|\cdot\|_{1,1}$ is the maximum l^1 norm of the columns, and that the weights matrices W_τ are diagonal, then the SC from Theorem 3 is

$$\sup_{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}} \left\|\left(D_\Gamma W_\Gamma\right)^+ g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}\right\|_1 < \alpha, \tag{63}$$

where $g_{\overline{\gamma}} \in D_{\overline{\Gamma}}$ and $w_{\overline{\gamma},\overline{\gamma}}$ is the corresponding *a priori* probability factor from the diagonal of matrix $W_{\overline{\Gamma}}$.

We define a new measure of the coherence of a dictionary where we introduce the idea of weighting the correlations among atoms with respect to the *a priori* information we have on f. Hence, we the following weighted cumulative coherence is defined:

$$\mu_1^w(m,f) \triangleq \max_{\Gamma|\ \|\Gamma\|_0 = m} \max_{\substack{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}^{\overline{\Gamma}}, \\ \overline{\gamma} \notin \Gamma}} \sum_{\gamma \in \Gamma} | < g_\gamma, g_{\overline{\gamma}} > | \cdot w_{\gamma,\gamma}^{\Gamma} \cdot w_{\overline{\gamma},\overline{\gamma}}^{\overline{\Gamma}}. \tag{64}$$

This new coherence measure is not independent of the data to approximate: it considers that all functions in the dictionary do not have equal probability to be used.

We can develop the SC in the same way as in App. A to obtain an upper bound based on μ_1^w, i.e.

$$\begin{aligned}\sup_{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}} \left\|\left(D_\Gamma W_\Gamma\right)^+ g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}\right\|_1 = \\ \sup_{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}} \left\|\left(\left(D_\Gamma W_\Gamma^T\right)^T \left(D_\Gamma W_\Gamma^T\right)\right)^{-1} \left(W_\Gamma D_\Gamma^T\right) g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}\right\|_1 \leq \\ \left\|\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right)^{-1}\right\|_{1,1} \cdot \sup_{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}} \left\|\left(W_\Gamma D_\Gamma^T\right) g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}\right\|_1.\end{aligned} \tag{65}$$

The first norm term can be upper bounded as in (54)-(59). However, due to the diagonal weight matrices W_τ, we can not define a matrix A composed only of the off diagonal elements. In fact, to obtain the matrix used in the series we have to add and subtract the identity matrix in the following way:

$$\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right)^{-1} = \left(I + A_w\right)^{-1} = \left(I + \left(\left(\left(W_\Gamma D_\Gamma^T\right)\left(W_\Gamma D_\Gamma^T\right)^T\right) - I\right)\right)^{-1}. \tag{66}$$

Thus,

$$\left\| \left(\left(W_\Gamma D_\Gamma^T \right) \left(W_\Gamma D_\Gamma^T \right)^T \right)^{-1} \right\|_{1,1} \leq \frac{1}{1 - \|A_w\|_{1,1}}. \tag{67}$$

The $1,1-norm$ of A_w can be expressed this time as:

$$\|A_w\|_{1,1} = \sup_{g_\gamma \cdot w_\gamma} \sum_{l \neq \gamma} | < g_l, g_\gamma > | \cdot w_l \cdot w_\gamma + |w_\gamma^2 - 1|, \tag{68}$$

where the summation comes from the off-diagonal elements ant the last term comes from the diagonal part. Notice that this term needs to converge to 0 for ensuring the convergence of the *Von Neumann* series as well as the bound in general. In the case of no information *a priori*, there are no weights and consequently the diagonal of A is always 0. This imposes that the information *a priori* about the original signal must be close to the truth i.e. atoms belonging to the set of "correct" should *not* be penalized. A wrong estimation of the *a priori* information to reduce the chances of "non-correct" atoms will lead the greedy algorithm to de-rail.

Under the assumption that "correct" atoms are relatively not (or lowly) penalized with respect to the rest, i.e. a good *reliable* guess of the data f is available, we may consider that the last term in (67) can be neglected, leading to:

$$\frac{1}{1 - \|A_w\|_{1,1}} \leq \frac{1}{1 - (\mu_1^w(m-1) + |w_\gamma^2 - 1|)} \approx \frac{1}{1 - \mu_1^w(m-1)}. \tag{69}$$

Coming back to Eq. (65), the second term can be bounded in the same manner as in Appendix A. Hence,

$$\sup_{g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}}} \left\| \left(W_\Gamma D_\Gamma^T \right) g_{\overline{\gamma}} \cdot w_{\overline{\gamma},\overline{\gamma}} \right\|_1 \leq \mu_1^w(m). \tag{70}$$

Finally, from (69) and (70) we obtain that if, and only if, the *a priori* information is *reliable* and

$$\frac{\mu_1^w(m)}{1 - \mu_1^w(m-1)} < \alpha \tag{71}$$

then, the recovery of the exact set of "correct" atoms will be accomplished by the greedy algorithm.

From this result it follows as well that since $\mu_1^w(m) \leq \mu_1(m)$, the consideration of a reliable *a priori* information can help a greedy algorithm with a dictionary that does not satisfy Theorem 1 to succeed in the recovery of the right set of functions.

■

Acknowledgments

The authors would like to thank Dr. Christophe de Vleeschouwer, Prof. Benoit Macq, Lorenzo Peotta and Rosa M. Figueras i Ventura for fruitful discussions and suggestions.

References

[1] O. Divorra Escoda and P. Vandergheynst, "Video coding using a deformation compensation algorithm based on adaptive matching pursuit image decompositions," in *IEEE International Conference on Image Processing (ICIP)*, Barcelona, September 2003.
[2] P. Frossard and P. Vandergheynst, "Efficient image representation by anisotropic refinement in matching pursuit," in *ICASSP*, vol. 3, Salt Lake City, May 2001.
[3] *Digital Video Processing and Communications.* Prentice Hall, 2001.
[4] E. J. Candès and D. L. Donoho, "Curvelets - a surprisingly effective non-adaptive representation for objects with edges." *Curves and Surfaces, L. L. S. et al., ed., Nashville, TN, (Vanderbilt University Press)*, pp. 123–143, 1999.
[5] M. N. Do, P. L. Dragotti, R. Shukla, and M. Vetterli, "On the compression of two-dimensional piecewise smooth functions," in *ICIP*, Thesalonica, October 2001.
[6] D. L. Donoho, "Wedgelets: Nearly-minimax estimation of edges," *Annals of Stat*, vol. 27, pp. 859–897, 1999.
[7] R. Figueras i Ventura, L. Granai, and P. Vandergheynst, "R-D analysis of adaptive edge representations," in *MMSP*, Virgin Islands, December 2002.
[8] P. V. R M Figueras i Ventura and P. Frossard, "Highly flexible image coding using non-linear representations," ITS, Tech. Rep., 2003.
[9] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Proc.*, vol. 41, no. 12, pp. 3397–3415, December 1993.
[10] P. Frossard, P. Vandergheynst, and R. M. Figueras i Ventura, "High flexibility scalable image coding," in *VCIP*. Lugano: SPIE, 2003.
[11] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," ICES, University of Texas at Austin, Austin, USA, Tech. Rep., 2003.
[12] ——, "Just relax: Convex programming methods for subset selection and sparse approximation," ICES, University of Texas at Austin, Austin, USA, Tech. Rep., 2004.
[13] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," IRISA, Rennes, France, Tech. Rep., 2004.
[14] M. Campani and A. Verri, "Motion analysis from first-order properties of optical," in *CVGIP: Image Understanding*, July 1992.
[15] Y. Altunbasak and A. M. Tekalp, "Closed-form connectivity-preserving solutions for motion compensation using 2-d meshes," *IEEE Trans. in Image Proc.*, vol. 6, no. 9, pp. 1255–1269, September 1997.
[16] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[17] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *IEEE Trans. in Image Proc.*, 2004.
[18] R. M. Figueras i Ventura, O. Divorra Escoda, and P. Vandergheynst, "A matching pursuit full search algorithm for image approximations," ITS-STI/EPFL, Tech. Rep. ITS-2004.031, December 2004.
[19] Y. Pati, R. Reziifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition."
[20] S. Martello and P. Toth, *Knapsack problems: algorithms and computer implementations.* New York: Wiley, 1990.
[21] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
[22] V. N. Temlyakov, "Weak greedy algorithms," Department of Mathematics, University of South Carolina, Columbia, Tech. Rep., 1999.
[23] A. Conn, N. Gould, and P. Toint, *Trust Region Methods.* http://www.fundp.ac.be/˜phtoint/pht/trbook.html: SIAM, 2000.
[24] Y. Lu, C. Lu, and Z. Li, "A modified space frequency decomposition algorithm for visual motion," in *ICME*, Baltimore, 2003.
[25] L. Peotta and P. Vandergheynst, "Mp in block quasi-incoherent dictionaries," Signal Processing Institute, EPFL, Lausanne, Switzerland, Tech. Rep., 2003.
[26] P. Nillius and J. O. Eklundh, "Fast block matching with normalized cross-correlation using walsh transforms," Computational Vision and Active Perception Laboratory (CVAP), KTH, Tech. Rep., 2002.
[27] *Probability, Random Variables, and Stochastic Processes*, 3rd ed. McGrawHill, 1991.
[28] M. P. Queluz, "Multiscale motion estimation and video compression," Ph.D. dissertation, Laboratoire de Telecommunications et Teledetection, UCL, Louvain la Neuve, Belgique, 1996.
[29] T. Aach, A. Kaup, and R. Mester, "Combined displacement estimation and segmentation of stereo image pairs based on gibbs random fields," in *ICASSP*, 1990.
[30] Kinderman and J. Snell, *Contemporary Mathematics: Markov Fields and Their Applications.* American Mathematical Society Providence, 1980, vol. 1.
[31] A. Rahmoune, P. Vandergheynst, and P. Frossard, "Mp3d: Highly scalable video coding scheme based on matching pursuit," in *ICASSP*, 2004.